(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 0 955 592 A2

# (12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
10.11.1999 Bulletin 1999/45

(51) Int Cl.[6]: G06F 17/30

(21) Application number: 99303432.1

(22) Date of filing: 30.04.1999

(84) Designated Contracting States:
AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE
Designated Extension States:
AL LT LV MK RO SI

(30) Priority: 07.05.1998 AU PP340598
07.05.1998 AU PP340898
07.05.1998 AU PP341098

(71) Applicant: CANON KABUSHIKI KAISHA
Tokyo (JP)

(72) Inventor: Yourlo, Zhenya
Roseville, New South Wales 2069 (AU)

(74) Representative:
Beresford, Keith Denis Lewis et al
BERESFORD & Co.
High Holborn
2-5 Warwick Court
London WC1R 5DJ (GB)

## (54) A system and method for querying a music database

(57) A system and method for querying a music database (302), the database containing a plurality of indexed pieces of music, where the query (104) is performed by forming a database request consisting of a conditional expression relating to the name and/or attributes of the desired piece of music. Associated parameters are derived from the database query, and compared with corresponding parameters for the other pieces of music in the database (302). A desired piece of music is determined by searching for a minimum distance between the database query parameters and those associated with the pieces of music in the database (302).

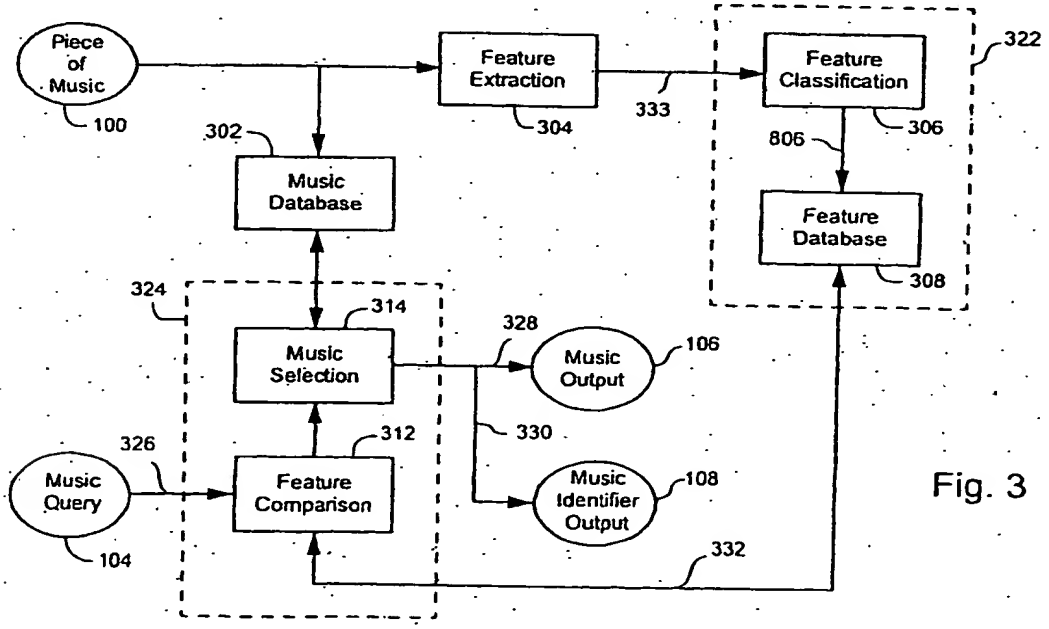


Fig. 3

EP 0 955 592 A2

Printed by Jouve, 75001 PARIS (FR)

Description

## FIELD OF THE INVENTION

[0001] The present invention relates to the field of music systems and, in particular, to the identification and retrieval of particular pieces of music or alternately, attributes of a desired piece of music, from a music database on the basis of a query composed of desired features and conditional statements.

## BACKGROUND OF THE INVENTION

[0002] Retrieval of music or music attributes from a database requires, in common with generic database functionality, a query method which is powerful and flexible, and preferably intuitively meaningful to the user. This in turn requires that the database contain music which has been classified in a manner which is conducive to systematic search and sort procedures. This latter aspect in turn requires that pieces of music be characterised in a manner which permits of such classification.

[0003] Thus a hierarchy of requirements or elements which make up a music database system are as follows:

- characterising music using attributes useful in a classification scheme
- classifying music in a meaningful searchable structure, and
- querying the database so formed, to yield meaningful results.

[0004] The hierarchy, has been defined "bottom up" since this presents a more meaningful progression by which the invention can be described.

[0005] When considering audio signals in general, and in particular those relating to music, the nature of the signals may be considered in terms of various attributes which are intuitively meaningful. These attributes include, among others, tempo, loudness, pitch and timbre. Timbre can be considered to be made up of a number of constituent sub-features including "sharpness" and "percussivity". These features can be extracted from music and are useful in characterising the music for a classification scheme.

[0006] The publication entitled "Using Bandpass and Comb Filters to Beat-track Digital Audio" by Eric D. Scheirer (MIT Media Laboratory, December 20, 1996) discloses a method for extraction of rhythm information or "beat track" from digital audio representing music. An "amplitude-modulated noise" signal is produced by processing a musical signal through a filter bank of bandpass filters. A similar operation is also performed on a white noise signal from a pseudo-random generator. Subsequently the amplitude of each band of the noise signal is modulated with the amplitude envelope of the corresponding band of the musical filter bank output. Finally the resulting amplitude modulated noise signals are summed together to form an output signal. It is claimed that the resulting noise signal has a rhythmic percept which is significantly the same as that of the original music signal. The method described can run in real-time on a very fast desktop workstation or alternately, a multi-processor architecture may be utilised. This method suffers from the disadvantage of being highly computationally intensive.

[0007] Percussivity is that attribute which relates to a family of musical instruments known as "percussion" when considering an orchestra or a band. This family includes such musical instruments as drums, cymbals, castanets and others. Processing of audio signals in general and musical signals in particular, benefits from the ability to estimate various attributes of the signals, and the present invention is concerned with estimating the attribute of percussivity.

[0008] A number of different methods have been used to estimate percussivity of a given signal, such methods including those broadly based upon:

- Short-time power analysis
- statistical analysis of signal amplitude
- comparison of harmonic spectral component power with total spectral power

[0009] Short-time signal power estimation involves calculation of an equivalent power (or an approximation thereof) within a short segment or "window" of a signal under consideration. The power estimate can be compared to a threshold in order to determine whether the portion of the signal within the window is percussive in nature. Alternatively, the power estimate can be compared to a sliding scale of thresholds, and the percussive content of the signal classified with reference to the range of thresholds.

[0010] Statistical analysis of signal amplitude is typically based upon a "running mean" or average signal amplitude value, where the mean is determined for a window which slides across the signal under consideration. By sliding the window, the running mean is determined over a pre-determined time period of interest. The mean value at each window position is compared to mean values for other windows in a neighborhood in order to determine whether signal variations

in the running mean are sufficiently large to signify that the signal is percussive.

**[0011]** Harmonic spectral component power analysis involves taking a windowed Fourier transform of the signal in question over the time period of interest, and then examining the resulting set of spectral components. The spectral components which are indicative of harmonic series are removed. It is noted that such harmonic series components typically represent local maxima in the overall spectral envelope of the signal. After removing the harmonic series spectral components, remaining spectral components substantially consist only of the inharmonic components of the signal, these being considered to represent percussive components of the signal. The total power in these inharmonic components is determined and compared with a total signal power for all components, harmonic and non-harmonic, to yield an indication of percussivity.

**[0012]** The aforementioned analysis methods are typically intended to identify a range of signal attributes, and thus suffer from relatively limited accuracy, and a tendency to produce false or unreliable percussivity estimates. The methods are also relatively complex and thus expensive to implement, particularly the harmonic spectral component estimation method.

**[0013]** U.S. Patent No. 5,616,876 (Cluts et al) entitled "System and Methods for Selecting Music on the Basis of Subjective Content" describes an interactive network providing music to subscribers which allows a subscriber to use a seed song to identify other songs similar to the seed song, the similarity between songs being based on the subjective content of the songs, as reflected in style tables prepared by editors. The system and methods described in this publication are based on the manual categorisation of music, with the attendant requirement for human participation in the process, with the resultant speed, accuracy and repeatability of the process limited by human attributes.

**[0014]** The publication entitled "Content - Based Classification, Search, and Retrieval of Audio" by Erling et al (IEEE Multimedia Vol. 3, No. 3, 1996, pp.27-36) discloses indexing and retrieving short audio files (i.e. "sounds") from a database. Features from the sound in question are extracted, and feature vectors based on statistical measures relating to the features are generated. Both the sound and the set of feature vectors are stored in a database for later search and retrieval. A method of feature comparison is used to determine whether or not a selected sound is similar to another sound stored in the database. The feature set selected does not include tempo and thus the system will not perform well in differentiating between pieces of music. Furthermore, the method determines features which provide scalar statistical measures over short time windows. Furthermore, the method uses features such as bandwidth which are not readily conceptualized in terms of impact of music selection.

**[0015]** It is seen from the above that existing arrangements have shortcomings in all elements in the hierarchy of requirements described, and it is an object of the invention to ameliorate one or more disadvantages of the prior art.

## SUMMARY OF THE INVENTION

**[0016]** According to one aspect of the invention, there is provided a method for querying a music database, which contains a plurality of pieces of music wherein the pieces are indexed according to a plurality of parameters, the method comprising the steps of:

forming a request which specifies one or more pieces of music and/or associated parameters and one or more conditional expressions;
determining associated parameters for the specified pieces of music if the parameters have not been specified;
comparing the specified parameters and corresponding parameters associated with other pieces of music in the database;
calculating a distance based on the comparisons;
identifying pieces of music which are at distances from the specified pieces of music as to satisfy the conditional expressions.

## BRIEF DESCRIPTION OF THE DRAWINGS

**[0017]** A preferred embodiment of the present invention will be described in detail with reference to the accompanying drawings in which:

Fig. 1 depicts a music database system in a kiosk embodiment;
Fig. 2 illustrates a music database system in a network embodiment;
Fig. 3 provides a functional description of a music database system;
Fig. 4 illustrates a generic feature extraction process;
Fig. 5 depicts the tempo feature extraction process;
Fig. 6 presents a further illustration of the tempo feature extraction process;
Fig. 7 depicts a process flow diagram for a preferred embodiment of the percussivity estimator;

Fig. 8 presents more detail of the preferred embodiment;
Fig. 9 illustrates a preferred embodiment of a comb filter;
Fig. 10 depicts a linear function obtained from the comb filter output energies;
Fig. 11 presents an accumulated histogram of a signal having an overall high percussivity;
Fig. 12 presents an accumulated histogram of a signal having an overall low percussivity;
Fig. 13 illustrates a typical percussive signal;
Fig. 14 depicts a generic feature classification process;
Fig. 15 shows a database query process - music identifiers supplied in query;
Fig. 16 illustrates a database query process - music features supplied in query;
Fig. 17 illustrates a distance metric used to assess similarity between two pieces of music; and
Fig. 18 - 21 depict feature representations for four pieces of music; and
Fig. 22 depicts a general purpose computer upon which the preferred embodiment of the invention can be practiced.

## DETAILED DESCRIPTION

[0018] Fig. 1 depicts a music database system in a kiosk 102 embodiment. For the purpose of the description it is noted that "kiosk" is a term of art denoting a public access data terminal for use in, for example, information data retrieval and audio output receipt. In this embodiment, the kiosk 102 owner/operator inputs pieces of music 100 into the kiosk 102 where they are classified and stored in a database for future retrieval. A music lover comes up to the kiosk and inputs a music query 104 to the kiosk 102, which after performing a search of the kiosk music database based on parameters in the music query 104, outputs a desired piece of music 106 which is based on the music query 104. The kiosk 102 also outputs music identifiers 108 associated with the desired piece of music 106. Such identifiers could include, for example, the name of the piece of music.

[0019] Fig. 2 illustrates a music database system in a network embodiment. In this embodiment, a plurality of music database servers 202 are connected to a network 206 via access lines 204. The Server 202 owner/operators input pieces of music 100 into the servers 202 where they are classified and stored in the database for future retrieval. Servers may be embodied in various forms including by means of general purpose computers as described in Fig. 4 below. A plurality of music database clients are also connected to the network 206 via access lines 208. A client owner inputs a music query 104 into the client 210, which establishes a connection to the music database server 202 via a network connection comprising access line 208, network 206 and access line 204. The server 202 performs a search of the music database based upon user query 104, and then outputs a desired piece of music 106, which is based on the music query 104, across the same network connection 204-206-208. The server 202 also outputs music identifiers 108 associated with the desired piece of music 106. Such identifiers could include, for example, the name of the piece of music.

[0020] Fig. 3 provides a functional description of the music database system. The database system performs two high level processes, namely (i) inputting pieces of music 100, classifying them and storing them in the database for later search and retrieval and (ii) servicing queries 104 to the music database system consequent upon which it outputs a desired piece of music 106 and/or music identifiers 108 associated with the desired piece of music 106. Such identifiers could include, for example, the name of the piece of music. Considering first the music input and classification process, a piece of music 100 is input, it then undergoes feature extraction 304 after which the features are classified 306 and stored in feature database 308. In parallel with this process, the actual music piece itself 100 is stored in music database 302. Thus the piece of music 100 and its associated representative features are stored in two databases 302 and 308. Next considering the database query process, the user query 104 is input whereupon feature comparison 312 is performed between the features associated with the user query 104 and the features of the pieces of music stored in the feature database 308. After a successful search, a music selection process 314 extracts the desired piece of music 106 from music database 302 on the basis of the feature comparison 312, and outputs the desired piece of music 106 and/or music identifiers 108 associated with the desired piece of music 106.

[0021] Fig. 4 depicts a generic feature extraction process. Recalling from the functional description of the database system in Fig. 3, the piece of music 100 is input, it then undergoes feature extraction 304 after which the features are classified 306 and stored in feature database 308. In Fig. 5, the piece of music 100 is input, and feature extraction process 304 is seen to include, in this illustration, four parallel processes, one for each feature. The tempo extraction process 402 operates upon the input piece of music 100 to produce tempo data output 404. The loudness extraction process 406 operates upon the input piece of music 100 to produce loudness data output 408. The pitch extraction process 410 operates upon the input piece of music 100 to produce pitch data output 412. The timbre extraction process 414 operates upon the input piece of music 100 to produce sharpness data output 416 and percussivity data output 418. Thus, referring again to Fig. 3 it is seen that for this example, the output line 332 between the feature comparison process 312 and the feature database 308 is handling four different data sets 504, 508, 512, 516.

[0022] Fig. 5 shows the tempo feature extraction process 402 (described in Fig. 4) and will be described in some

detail. Tempo extraction firstly involves determination of the onset signal 620 from the piece of music 100, and then filtering this onset signal through a bank of comb filters. Finally the energy in the comb filters accumulated over substantially the entire duration of the piece of music 100 provides the raw tempo data 404 indicative of the tempo or tempi (various tempos) present in the piece of music 100 substantially over it's duration 602. This set of processes is preferably implemented in software. Alternatively, a number of processes and sub-processes can, if desired, be performed, for example, on the audio input card 216, where say a Fast Fourier Transform (FFT) can be performed using a Digital Signal Processor (DSP). Furthermore, comb filters, described in relation to feature extraction, can also be implemented using a DSP on the audio card 216. Alternatively, these processes can be performed by the general purpose processor 102. In Fig. 5, an input music signal 100 is partitioned into windows 502 and the Fourier coefficients in each window determined 504. This is an expanded view of the Fast Fourier Transform process 522. After calculating the FFT, the coefficients in each window or "bin" are summed 506 and the resulting signal 524 is low pass filtered 508, then differentiated 510, and finally half wave rectified 512 to produce onset signal 618 (see Fig. 6 also).

**[0023]** Turning to Fig. 6, a waveform representation of the process described in Fig. 5 is shown. After windowing the input music signal 100, the signal in each time window 604 is processed by a Fast Fourier Transform (FIT) process to form an output signal 620 which is shown pictorially as frequency components 606 in frequency bins 622-624 which are divided into individual time windows 604. Output signal 620 then has its frequency component amplitudes 606 in the various frequency bins 622-624 added by an addition process 608. This summed signal, which may be considered as an energy signal, has positive polarity and is passed through a low pass filter process 610, whose output signal 628 is differentiated 612 to detect peaks, and then half-wave rectified 614 to remove negative peaks, finally producing onset signal 618. The music signal is processed across substantially the full duration of the piece of music 100. In an alternate embodiment, the onset signal 618 could be derived by sampling signal 628, comparing consecutive samples to detect positive peaks of the signal 614, and generating pulses 628 each time such a peak is detected. A brief explanation about the effect of partitioning the signal into time windows is in order. Summing frequency component amplitudes in each window is a form of decimation (i.e. reduction of the sampling frequency), since the number of digitised music samples in a window are summed to form one resultant point. Thus selection of the window size has the effect of reducing the number of sample points. The optimum selection of window size requires a balance between the accuracy of the resultant representation of the feature, and compression of the data in order to reduce computational burden. The inventor has found that a 256 point FFT (equivalent to an 11.6 msec music window size) yields good performance when using the resultant feature for comparing and selecting music pieces in regard to tempo. Once significant changes in the spectrum (i.e. the starting points of notes 616) are located, the onset signal 618 is passed through a bank of comb filters in order to determine the tempo. As noted previously, comb filters can be implemented using DSPs on the audio card 116 or alternatively by using general purpose processor 102. Each comb filter has a transfer function of the form:

$$y_t = \alpha y_{t-\tau} + (1 - \alpha) x_t$$

where:

- $y_t$ represents the instantaneous comb filter output
- $y_{t-\tau}$ represents a time delayed version of the comb filter output
- $x_t$ represents the onset signal (618).

**[0024]** Each of these comb filters has a resonant frequency (at which the output is reinforced) determined by the parameter $1/\tau$. The parameter $\alpha$ (alpha) corresponds to the amount of weighting placed on previous inputs relative to the amount of weighting placed on current and future inputs. The onset signal 618 is filtered through the bank of comb filters, whose resonant frequencies are placed at frequencies which are at multiple sample spacings resulting from windowing. The filters should typically cover the range from about 0.1Hz through to about 8Hz. The filter with the highest energy output at each sample point is considered to have "won", and a tally of wins is maintained for each filter in the filterbank, for example by using a power comparator to determine the highest energy and a counter to tally the "wins". After the onset signal 618 over substantially the full duration 602 of the piece of music 100 has been filtered, the filter that has the greatest tally is said to be the dominant tempo present in the original music signal 100. Secondary tempo's may also be identified using the method.

**[0025]** The timbre of a sequence of sound, which feature is characteristic of the difference in sounds say between two musical instruments, is largely dependent upon the frequencies present, and their respective magnitudes.

**[0026]** The spectral centroid provides an estimate of "brightness" or "sharpness" of the sound, and is one of the metrics used in the present embodiment in relation to extraction of timbre. This brightness characteristic is given by:

$$S = \sum_{w} (f)(A) / \sum A$$

where:

S = spectral centroid
*f* = frequency
A = Amplitude
W = Window selected

**[0027]** In order to differentiate between the timbral characteristics of different audio signals, the present embodiment makes use of the Fourier transform of successive 0.5 second windows of the audio signal 100 in question. There is no necessary relationship between the window size used for loudness feature extraction and that used for tempo or other feature extraction. Other techniques for extracting timbre may be used.

**[0028]** Percussivity is that attribute which relates to a family of musical instruments known as "percussion" when considering an orchestra or a band. This family includes such musical instruments as drums, cymbals, castanets and others.

**[0029]** Fig. 7 depicts a flow diagram of a preferred embodiment of the percussivity estimator disclosed in the present invention. An input signal 736 on line 700 is analysed for percussivity during a time interval of interest 742. The input signal 736 is described in an inset 702 where the signal 736 is depicted on axes of time 706 and amplitude 704. The signal 736 is operated upon by a windowing process 710 which outputs a windowed signal on line 734, the windowed signal being shown in more detail in an inset 712. In the inset 712 windows, exemplified by a window 738, each having a predetermined width 708, are overlapped with each other to an extent 776. Each window 738 is passed through a bank of comb filters 740 which is made up of individual comb filters exemplified by comb filter 718. The structure and operation of an embodiment of the comb filter 718 is presented in more detail in relation to Fig. 3. The comb filter 718 integrates the energy of the signal 736 across the particular window 738 being considered. The bank of comb filters 740, outputs a peak energy 726 for each comb filter 718 in the bank of comb filters 740 for the window 738 being considered, representing the energy at frequencies corresponding to the comb filter. This is shown in an inset 724. It is noted that the outputs exemplified by output 726 of the comb filter bank 740 are represented on axes of amplitude and frequency, and are spaced according to the frequencies corresponding to the individual comb filters 718. The output from the comb filter bank 740 on line 720 is processed by a gradient process 722 which determines a straight line of best fit 732 which approximates the output signal exemplified by signal 726.

**[0030]** Fig. 8 presents a more detailed description of the preferred embodiment of the percussivity estimator as it relates to a digitised input signal. Given an input signal on line 800 to be analysed, the signal is first digitised in process 802. The digitised signal which is then output on line 804 is windowed by process 806 into 100 msec windows, with a 50% overlap between windows. Each window is passed through a bank of comb filters 740 represented by process 810. The comb filters making up process 810 are spaced at frequencies between 200 Hz and 3000 Hz. The number and spacing of the individual comb filters 718 in the comb filter bank are discussed in more detail in relation to Fig. 9. The linear function on line 812 which is formed from the peak energy output of each comb filter comprising the comb filter bank process 810 is passed to a gradient process 814. The gradient process 814 determines a straight line of best fit which approximates the linear function which is output by the comb filter process 810 on line 812, and outputs the straight line function on line 816 for further processing.

**[0031]** Fig. 9 depicts a block diagram of a preferred embodiment of an individual comb filter 718 used in the embodiment of the percussivity estimator. The comb filter 718 is used as a building block to implement the bank of comb filters 740 (see Fig. 7). As described in relation to Fig. 6, each comb filter 718 has a time response which can be represented mathematically as follows:

$$y(t) = a^{*}y(t-T) + [1-a]^{*}x(t) \qquad [1]$$

where:

x(t) is an input signal 900 to the comb filter;
y(t) is an output signal 906 from the comb filter;
T is a delay parameter determining the period of the comb filter; and
a is a gain factor determining the frequency selectivity of the comb filter.

[0032] For each comb filter 718 in the bank of comb filters 740 (see Fig. 7), the delay factor T is selected to be an integral number of samples long, the sample attributes being determined by process 802 (see Fig. 8). In the preferred embodiment of the comb filter bank 740, the number of filters 718 in the bank 740 is determined by the number of integral sample lengths between the resonant frequency edges, these edges being defined in the embodiment described in relation to Fig. 8 to be 200 Hz and 3000 Hz. The individual filters 718 need not be equally spaced between the frequency edges, however they need to provide substantial coverage of the entire frequency band between the edges.

[0033] Fig. 10 depicts a linear function 1000 which is formed from the peak energy outputs of each comb filter 718 in the comb filter bank 740. The ordinate 402 represents the peak energy output 126 of each comb filter 118 in the filterbank 1040, while the abscissa 1004 represents the resonant frequency of each filter 718. Thus exemplary point 1012 indicates that a filter having a resonant frequency 1008 has output a peak energy output 1010 for the particular window being considered. A line of best fit 1006 is shown, having a gradient 1014 which is representative of the percussivity of the signal 736 within the particular window in question.

[0034] Fig. 11 depicts how an aggregate of individual gradients, say 1014, each having been determined for a particular window, say 738, can be consolidated and represented in the form of a histogram 1100 covering the entire period of interest 742 for the signal 736 being considered. An ordinate 1102 represents the fraction of time during the time interval 742 during which a particular percussivity is found to be present. An abscissa 1104 represents a normalised percussivity measure, which can be determined by normalising all measured percussivity values during the time interval of interest 742 by the maximum percussivity value during the period 742. Thus, point 1106 indicates that a normalised percussivity value 1110 is found to be present for a fraction 1108 of the total time 742. It is noted that the area under the curve 1100 can also be normalised for different signals being analyzed, in order to enable comparison of percussivity between the different signals. Fig. 11 represents a histogram for a signal having an overall high percussivity.

[0035] Fig. 12 depicts a percussivity histogram for a different signal than the one considered in Fig. 11, the signal in Fig. 12 having an overall low percussivity.

[0036] Fig. 13 depicts a typical percussive signal 1304 in the time domain, where the signal 1304 is plotted as a function of an amplitude axis 1300 and a time axis 1302.

[0037] The loudness feature is representative of the loudness over substantially the full duration of the piece of music 100 (see Fig. 1). The piece of music 100 is first partitioned into a sequence of time windows, which for the purpose of classification and comparison on the basis of loudness, should be preferably about one half a second wide. There is no necessary relationship between the window size used for loudness feature extraction and that used for tempo or other feature extraction. The Fourier transform of the signal in each window is taken, and then the power in each window is calculated. The magnitude of this power value is an estimate of the loudness of the music within the corresponding half-second interval. Other methods of extracting loudness are known.

[0038] Pitch is another feature in the present embodiment determined by the feature extraction means in order to represent music while storing a new piece of music into the music database. The localised pitch is determined over a small window (say 0.1 seconds in this instance) by using a bank of comb filters. There is no necessary relationship between the window size used for pitch feature extraction and that used for tempo or other feature extraction. These comb filters have resonant frequencies covering a range of valid pitches. Advantageously this includes frequencies from around 200Hz up to around 3500Hz, and the filters are spaced at intervals determined by the rate at which the original musical signal was sampled. The sampled signal is filtered through the filter bank, and the comb filter that has the greatest output power will have a resonant frequency corresponding to the dominant pitch over the window in question. From these resulting pitches, a histogram of dominant pitches present in the original music is formed. This procedure is followed over substantially the entire duration of the piece of music. The method of pitch extraction employed is one of a number of methods for pitch extraction which exists and other methods may be used.

[0039] Returning to Fig. 3, and considering the music input and classification process, when the piece of music 100 is input, it then undergoes feature extraction 304 after which the features are classified 306 and stored in feature database 308. Substantially in parallel with this process, the actual music piece itself 100 is stored in music database 302. Thus the piece of music 100 and the associated representative features are stored in two distinct but related databases 302 and 308 respectively. If the music is initially derived from an analogue source, it is first digitised before being input into the feature extraction process 304. The digitisation step may be implemented by way of a standard soundcard or, if the music is already in digital form, this digitisation step may be bypassed and the digital music used directly for 100. Thus, arbitrary digitization structures including the Musical Instrument Digital Interface (MIDI) format and others may be supported in the system. There are no specific requirements in terms of sampling rate, bits per sample, or channels, but it should be noted that if higher reproduction quality is desirable it is preferable to select an audio resolution close to that of a CD.

[0040] Fig. 14 depicts a generic feature classification process. Extracted feature signals 404, 408, 412, 416 and 418 (refer Fig. 4) are accumulated in process step 1404 as histograms over substantially the whole duration of the piece of music 100 resulting in an indicative feature output 1406 for each extracted feature signal. This output 1406 is stored

in the feature database 308. By identifying the N highest tempo's in the manner described in Figs. 5 and 6, a histogram describing the relative occurrence of each tempo across substantially the whole duration of the piece of music 100 can be formed. Similarly, by identifying the M highest volumes, a histogram describing the relative occurrence of each loudness across substantially the whole duration of the piece of music 100 can be formed. Again, by identifying the K dominant pitches, a histogram describing the relative occurrence of each pitch across substantially the whole duration of the piece of music 100 can be formed. The spectral centroid is advantageously used to describe the sharpness in a window. This can be accumulated as a histogram over substantially the full duration of the piece of music being analyzed and by identifying P sharpnesses (one for each window), a histogram describing the relative occurrence of each sharpness across substantially the whole duration of the piece of music 100 can be formed. Accumulation of features as histograms across substantially the entire duration of pieces of music yields a duration independent mechanism for feature classification suitable for search and comparison between pieces of music. This forms the foundation for classification in the music database system. The spectral centroid is advantageously used to describe the percussivity in a window. This can be accumulated as a histogram over substantially the full duration of the piece of music being analyzed and by identifying P percussivities (one for each window), a histogram describing the relative occurrence of each percussivity across substantially the whole duration of the piece of music 100 can be formed.

**[0041]** Fig. 15 describes a database query process where music identifiers are supplied in the query. A music query 104 (see Fig. 1) may take on a number of forms which include, but are not limited to:

(1) a set of names of known pieces of music and a degree of similarity/dissimilarity specified by a conditional expression (shown underlined) for each piece (e.g. very much like "You can hear me in the harmony" by Harry Conick Jr., a little like "1812 Overture" by Tchaikovsky, and not at all like "Breathless" by Kenny G);
(2) a set of user specified features and a similarity/dissimilarity specification in the form of a conditional expression (e.g. something that has a tempo of around 120 beats per minute, and is mostly loud).

**[0042]** In Fig. 15, the music query 104, containing music identifiers and conditional expressions is input into the feature comparison process 312 (see Fig. 3). This process 312 includes the feature retrieval process 1502 which retrieves the features associated with the pieces of music named in the music query 104 from feature database 308. Next these retrieved features are passed to similarity comparison process 1504 which searches the feature database 308 for features satisfying the conditional expression contained in music query 104 as applied to the features associated with pieces of music named in music query 104. The results of this comparison are passed to the identifier retrieval process 1506 which retrieves the music identifiers of the pieces of music whose features satisfy the conditional expressions as applied to the identifiers specified in music query 104. These identifiers are passed to the music selection process 314 which enables the output of the desired music 106 and/or music identifiers 108 from music database 302 and feature database 308 respectively.

**[0043]** Fig. 16 describes a database query process where music features are supplied in the music query 104. The music query 104, containing music features and conditional expressions, is available at the query stage 104 and thus in this case the feature retrieval process 1502 is bypassed (see Fig. 15). Next these provided features are passed to the similarity comparison process 1604 which searches the feature database 308 for features satisfying the conditional expression contained in music query 104 as applied to the features provided in the music query 104. The results of this comparison are passed to the identifier retrieval process 1606 which retrieves the music identifiers of the pieces of music whose features satisfy the conditional expressions in relation to the identifiers specified in music query 104. These identifiers are passed to the music selection process 314 which ensures the output of the desired music 106 and/or music identifiers 108 from music database 302 and feature database 308 respectively.

**[0044]** Considering the process of feature comparison 312, a similarity comparison is performed between the features of music stored by the system in the feature database 308 which correspond to pieces of music 100 stored in music database 302, and features corresponding to the music query 104. Since a number of different features (and feature representations) exist in the feature database 308, the comparisons between corresponding features are advantageously performed differently for each feature, for example:

- comparison between loudness features stored as histograms are made through the use of a histogram difference, or comparison of a number of moments about the mean of each histogram, or other methods that achieve the same goal;
- comparison between tempo features stored as histograms are accomplished by methods such as histogram difference, or comparison of a number of moments about the mean of each histogram or other methods that achieve the same goal,
- comparison between pitch features stored as histograms are performed using a histogram difference, or a comparison of a number of moments about the mean of each histogram. Other methods for comparison of pitch features may also be used.

- comparison between sharpness features stored as histograms are achieved through the use of methods such as histogram difference, or comparison of a number of moments about the mean of each histogram, or other methods that achieve the same goal, and
- comparison between percussivity features stored as histograms are achieved through the use of methods such as histogram difference, or comparison of a number of moments about the mean of each histogram, or other methods that achieve the same goal.

**[0045]** Once the comparison of each of the relevant features has been made, the overall degree of similarity is ascertained. A simple, yet effective way of determining this is through the use of a distance metric (also known as the Minkowski metric with $r = 1$), with each of the feature comparison results representing an individual difference along an orthogonal axis.

**[0046]** Fig. 17 illustrates a distance metric used to assess the similarity between two pieces of music where D is the distance between the two pieces of music 1708 and 1710 (only 3 features are shown for ease of representation). In this case, a smaller value of D represents a greater similarity. D is advantageously represented by:

$$\text{SQRT}((\text{loudness histogram difference})^2 + (\text{tempo histogram (difference)})^2 + (\text{pitch histogram difference})^2 + (\text{timbre histogram difference})^2)$$

**[0047]** Fig. 17 illustrates the distance between two pieces of music 1708, 1710, these pieces of music being defined in terms of three exemplary features namely pitch 1702, tempo 1704, and sharpness 1706. Distance D 1712 represents the distance between the pieces of music 1710 and 1708 when measured in this context.

**[0048]** The above method will be partially described for a specific query 104 namely "Find a piece of music similar to piece A", where the database contains pieces of music A, B, C, and D. This query 104 is of a type described in Fig. 15 where music identifiers (ie the name of the piece of music "A") and a conditional expression ("similar to") is provided in the query 104.

**[0049]** Each piece of music stored in the database is represented by a number of features that have been extracted when the pieces were classified and stored in the database. For the sake of simplicity the example presented is restricted to two features, namely tempo and sharpness, where both features are represented by simplified histograms.

**[0050]** The four music pieces to be considered are named A, B, C and D. Their corresponding feature histograms are illustrated in Figs. 18-21.

**[0051]** Fig. 18 illustrates a tempo histogram and a timbre (alternatively called sharpness) histogram for piece of music A. This piece of music is shown to have a tempo of 1 Hz (or 60 beats/min) (1800) for 0.5 or 50% of the time (1808) and a beat of 2 Hz (or 120 beats/minute) (1802) for 50% of the time (1808). The piece of music displays a brightness of 22050 Hz (1804) for 20% of the time (1810) and a brightness of 44100 Hz (1806) for 80% of the time (1812). Figs. 19 - 21 display similar features for pieces of music B - D.

**[0052]** When the query is presented, the following sequence of operations is performed:

- Comparison of the features of A and B
- Comparison of the features of A and C
- Comparison of the features of A and D
- Selection of the music that is least distant from A

**[0053]** Since all features of the music in the database are preferably represented as histograms, comparisons between these features is based on a comparison between the histograms. Two methods that are useful in forming this comparison are the histogram difference, and the comparison of moments.

- Considering the first method, the histogram difference is performed by comparing the relative occurrence frequencies of the different observations, taking the sum over all these comparisons and then normalising by the number of histograms being compared. If both histograms are individually normalised such that their individual integral sums are equal to 1.0, then the maximum histogram difference will be 2.0 (and if the absolute value of each comparison is taken, the minimum difference will be 0.0).

**[0054]** Considering the second method, comparison of moments is achieved by considering the differences between a number of moments about the origin of each histogram. The general form may be used to calculate moments about the origin:

$$\mu_k = \sum_{all\ x} x^k . f(x)$$

where:

$\mu_k$ is the Kth moment about the origin
$x^k$ is the Xth component of the histogram
$f(x)$ is the value of the histogram of $x^k$.

[0055] It is also common to normalise the moments with respect to the second moment about the origin, in order to make them independent of the scale of measurement:

$\mu_k \mu_2^{-k/2}$

[0056] With reference to Figs. 18 and 19, for the query 104 "similar to A" employing histogram difference, the calculation of distance is performed as follows:

[0057] The difference between A and B in regard to tempo is:

$$\frac{|0.5-0.33|+|0.5-0.33|+|0-0.33|}{2}=0.33$$

where the number of terms in the numerator is determined by the number of histogram points being compared, and the denominator is determined by the fact that two histograms are being compared.

[0058] Similarly, for A and B in regard to timbre:

$$\frac{|0.2-0.9|+|0.8-0.1|}{2}=0.7$$

[0059] Thus, distance between A and B is given by:

$$\sqrt{0.7^2+0.335^2}=0.776$$

[0060] If we consider the histograms in Figs. 18-21 for the features extracted from the piece of music A, B, C and D:

[0061] Music A, tempo histogram:

$$\mu_2 = 0.5 \times 1.0^2 + 0.5 \times 2.0^2 + 0 \times 3.0^2 = 2.50$$

$$\mu_3 = 0.5 \times 1.0^3 + 0.5 \times 2.0^3 + 0 \times 3.0^3 = 4.50$$

$$\mu_4 = 0.5 \times 1.0^4 + 0.5 \times 2.0^4 + 0 \times 3.0^4 = 8.50$$

$$\mu_3 \mu_2^{-3/2} = 1.14$$

$$\mu_4 \mu_2^{-4/2} = 1.36$$

[0062] Music A, sharpness histogram:

$$\mu_2 = 1.653 \times 10^9$$

$$\mu_3 = 7.076 \times 10^{13}$$

$$\mu_4 = 3.073 \times 10^{18}$$

$$\mu_3\mu_2^{-3/2} = 1.05$$

$$\mu_4\mu_2^{-4/2} = 1.12.$$

[0063] Music B, tempo histogram:

$$\mu_2 = 4.62$$

$$\mu_3 = 11.88$$

$$\mu_4 = 32.34$$

$$\mu_3\mu_2^{-3/2} = 1.20$$

$$\mu_4\mu_2^{-4/2} = 1.52$$

[0064] Music B, sharpness histogram:

$$\mu_2 = 6.321 \times 10^{8}$$

$$\mu_3 = 1.823 \times 10^{13}$$

$$\mu_4 = 5.91 \times 10^{17}$$

$$\mu_3\mu_2^{-3/2} = 1.15$$

$$\mu_4\mu_2^{-4/2} = 1.48$$

[0065] Comparisons for the query "similar to A":

[0066] A and B tempo:

$$|1.14 - 1.20| + |1.36 - 1.52| = 0.22$$

[0067] A and B sharpness:

$$|1.05 - 1.15| + |1.12 - 1.48| = 0.46$$

[0068] A and B distance:

$$\sqrt{0.22^2 + 0.46^2} = 0.5$$

[0069] The above analysis is only partially shown for the sake of brevity, however if fully expanded, it is seen that in both histogram difference and moment methods, Music B would be selected by the query 104 as being "similar to A", since the calculated distance between Music A and Music B is smallest when compared to C, D.

[0070] In the above example, the query 104 was "find a piece of music similar to piece A" and thus the method sought to establish which pieces of music B, C and D are at a distance smallest from A.

[0071] In a more complex query 104 of the form, for example, "find a piece of music very similar to A, a little bit like B, and not at all like C", the same general form of analysis as illustrated in the previous example would be used. In this case however, the other pieces of music in the database namely D, E, ..., K, ... would be assessed in order to establish which piece of music had features which could simultaneously satisfy the requirements of being at a minimum distance from A, a larger distance from B, and a maximum distance from C.

[0072] It is further possible to apply a weighting to each individual feature in order to bias the overall distance metric in some fashion (for example biasing in favour of tempo similarity rather than loudness similarity).

[0073] In considering similarity assessment on the basis of either the histogram difference, or the comparison of moments, these being applied to the attributes of pitch, loudness, tempo, and timbre (i.e. sharpness and percussivity), it is found that two-pass assessment provides better classification results in some cases. The two-pass assessment process performs a first assessment on the basis of loudness, percussivity and sharpness, and then a second sorting process based on tempo. In the present embodiments, it is found that the feature of pitch may be omitted from the similarity assessment process without significantly degrading the overall similarity assessment results.

[0074] In considering similarity assessment using the comparison of moments process, good results are produced by selecting particular moments for each feature as shown in the following table:

| Feature | Moments |
|---|---|
| loudness | mode, mean, variance |
| sharpness | mode, mean, variance |
| percussivity | variance |
| tempo | mode, mode tally, variance |

where "mean" and "variance" are determined in accordance with the following general form which expresses moments about the mean:

$$\mu_k = \sum_{all\ x} (x - \bar{x})^k \bullet f(x)$$

where:

"mean" = $\mu_k$ for k=1; and
"variance" = $\mu_k$ for k=2;

[0075] The "mode", in particular having regard to tempo, represents the most frequently occurring i.e. the "dominant" tempo in the tempo histogram, and is thus the tempo associated with the peak of the histogram. The "mode tally" is the amplitude of the peak, and represents the relative strength of the dominant tempo.

[0076] Application of clustering techniques to a complete set of moments corresponding to the extracted features, including the mode of each histogram, provides better classification results in some cases. Use of Bayesian estimation produces a "best" set of classes by which a given dataset may be classified.

[0077] Fig. 22 shows how the system can preferably be practised using a conventional general-purpose computer 2200 wherein the various processes described may be implemented as software executing on the computer 2200. In particular, the various process steps are effected by instructions in the software that are carried out by the computer 2200. The software may be stored in a computer readable medium, is loaded onto the computer 2200 from the medium, and then executed by the computer 2200. The use of the computer program product in the computer preferably effects an apparatus for (i) extracting one or more features from a music signal, said features including, for instance, tempo, loudness, pitch, and timbre, (ii) classification of music using extracted features, and (iii) method of querying a music database. Corresponding systems upon which the above method steps may be practised may be implemented as described by software executing on the above mentioned general-purpose computer 2200. The computer system 2200 includes a computer module 2202, an audio input card 2216, and input devices 2218, 2220. In addition, the computer

system 2200 can have any of a number of other output devices including an audio output card 2210 and output display 2224. The computer system 2200 can be connected to one or more other computers using an appropriate communication channel such as a modem communications path, a computer network, or the like. The computer network may include a local area network (LAN), a wide area network (WAN), an Intranet, and/or Internet. Thus, for example, pieces of music 100 may be input via audio input card 2216, music queries may be input via keyboard 2218, desired music 106 may be output via audio output card 2210 and desired music identifiers such as the names of the desired pieces of music may be output via display device 2224. The network embodiment shown in Fig. 2 would be implemented by using the communication channels to connect server computers to the network 206 via access lines 204. Client computers would be connected to the network 206 via access lines 208 also using the computer communication channels. The computer 2202 itself includes a central processing unit(s) (simply referred to as a processor hereinafter) 2204, a memory 2206 which may include random access memory (RAM) and read-only memory (ROM), an input/output (IO) interface 2208, an audio input interface 2222, and one or more storage devices generally represented by a block 2212. The storage device(s) 2212 can include one or more of the following: a floppy disk, a hard disk drive, a magneto-optical disk drive, CD-ROM, magnetic tape or any other of a number of non-volatile storage devices well known to those skilled in the art. Each of the components 2204, 2206, 2208, 2212 and 2222, is typically connected to one or more of the other devices via a bus 2204 that in turn can include data, address, and control buses. The audio interface 2222 is connected to the audio input 2216 and audio output 2210 cards, and provides audio input from the audio input card 2216 to the computer 2202 and from the computer 2202 to the audio output card 2210.

**[0078]** The preferred embodiment of the invention can, alternatively, be implemented on one or more integrated circuits. This mode of implementation allows incorporation of the various system elements into individual pieces of equipment where the functionality provided by such apparatus is required.

**[0079]** In the case where software is provided to configure a general purpose computer 2200 to operate in accordance with the present invention, the software may be modulated onto a carrier signal and downloaded through a computer network, such as the internet, to the general purpose computer 2200.

**[0080]** The foregoing describes a number of embodiments for the present invention. Further modifications can be made thereto without departing from the scope of the inventive concept.

## Claims

1. A method for querying a music database, which contains a plurality of pieces of music wherein the pieces are indexed according to one or more parameters, the method comprising the steps of:

   (a) forming a request which specifies one or more pieces of music and/or associated parameters and one or more conditional expressions;
   (b) determining associated parameters for the specified pieces of music if the parameters have not been specified;
   (c) comparing the specified parameters and corresponding parameters associated with other pieces of music in the database;
   (d) calculating a distance based on the comparisons; and
   (e) identifying pieces of music which are at distances from the specified pieces of music as to satisfy the conditional expressions.

2. A method according to claim 1, including the further steps of:

   (f) outputting at least one of (fa) the identified pieces of music and (fb) the names of said pieces.

3. A method according to claim 1 or 2, whereby calculating a distance in step (d) comprises the sub-steps of:

   (da) calculating a distance based on comparison of a loudness, a percussivity, and a sharpness; and subsequently
   (db) sorting on the basis of a tempo.

4. A method according to any preceding claim, whereby the comparing in step (c) is with pieces of music in a class of the plurality of pieces of music in the database.

5. A method according to claim 2, whereby the at least one of (fc) identified pieces of music and (fd) the names of said pieces are in a class of the plurality of pieces of music in the database.

6. A method according to any preceding claim, whereby a classification according to which the pieces of music are indexed uses feature extraction, the method further comprising the steps of:

(g) segmenting a piece of music over time into a plurality of windows;
(h) extracting one or more features in each of said windows; and
(i) arranging the features in histograms wherein the histograms are representative of the features over the entire piece of music.

7. A method according to claim 6, whereby a first feature extracted in step (h) is at least one tempo extracted from a digitised music signal, the feature extraction comprising the further sub-steps of:

(ha) segmenting the music signal into a plurality of windows;
(hb) determining values indicative of the energy in each window;
(hc) locating the peaks of an energy signal which signal is derived from the energy values in each window;
(hd) generating an onset signal comprising pulses, where pulse peaks substantially coincide with the peaks of the energy signal;
(he) filtering the onset signal through a plurality of comb filter processes with resonant frequencies located according to frequencies derived from the window segmentation;
(hf) accumulating an energy in each filter process over a duration of the music signal; and
(hg) identifying the filter processes having Nth highest energies wherein resonant frequencies of the identified processes are representative of at least one tempo in the music signal.

8. A method according to claim 7, whereby the determination of the energy signal in sub-step (hb) comprises the further sub-sub-steps of:

(hba) determining transform components for the music signal in each window; and
(hbb) adding amplitudes of the components in each window to form a component sum, said component sum being indicative of energy in a window.

9. A method according to claim 7 or 8, whereby after locating the peaks of an energy signal in sub-step (hc) and prior to forming the onset signal generated in sub-step (hd) the method comprises the further sub-sub-steps of:

(hca) low pass filtering the energy signal.

10. A method according to claim 7, 8 or 9, whereby the onset signal is formed in sub-step (hd) according to the further sub-sub-steps steps of:

(hda) differentiating the energy signal; and
(hdb) half-wave rectifying the differentiated signal to form the onset signal.

11. A method according to claim 7, 8 or 9, whereby the onset signal is formed in sub-step (hd) according to the further sub-sub-steps of:

(hdc) sampling the energy signal;
(hdd) comparing consecutive samples to determine a positive peak; and
(hde) generating a single pulse when each positive peak is detected.

12. A method according to any of claims 7 to 11, wherein the filter process resonant frequencies span a frequency range substantially between 1Hz and 4Hz.

13. A method according to claim 6, whereby a second feature extracted in step (h) is a percussivity of a signal, the method comprising the sub-steps of:

(hh) segmenting the signal into a plurality of windows, and for each window;
(hi) filtering by a plurality of filters;
(hj) determining an output for each filter;
(hk) determining a function of the filter output values;
(hl) determining a gradient for the linear function; and

(hm) determining a percussivity as a function of the gradient.

14. A method according to claim 13, whereby the segmentation sub-step (hh) comprises the further sub-sub-steps of:

(hha) selecting a window width;
(hhb) selecting a window overlap extent; and
(hhc) segmenting the signal into windows each window having the selected window width and the windows overlapping each other to the selected overlap extent.

15. A method according to claim 13 or 14, whereby the filtering sub-step (hi) utilises comb filters.

16. A method according to claim 13, 14 or 15, whereby the gradient determination step (hl) is performed by determining a straight line of best fit to the linear function.

17. A method according to any of claims 13 to 16, whereby the percussivity values determined in step (hm) for a window for each window are consolidated in a histogram.

18. An apparatus for querying a music database, which contains a plurality of pieces of music wherein the pieces are indexed according to one or more parameters the apparatus comprising:

(a) a request means for forming a request which specifies one or more pieces of music and/or associated parameters and one or more conditional expressions;
(b) a parameter determination means for determining associated parameters for the specified pieces of music if the parameters have not been specified;
(c) a comparison means for comparing the specified parameters and corresponding parameters associated with other pieces of music in the database;
(d) a distance determination means for calculating a distance based on the comparisons; and
(e) a determination means for identifying pieces of music which are at distances from the specified pieces of music as to satisfy the conditional expressions.

19. An apparatus according to claim 18, wherein the apparatus further comprises:

(f) an output means for outputting the identified pieces of music and/or the names of said pieces.

20. An apparatus according to claim 18 or 19, whereby the distance determination means (d) comprises:

(da) a distance determination means for a loudness, a percussivity, and a sharpness; and
(db) a sorting means for a tempo.

21. An apparatus according to claim 18, 19 or 20, whereby the apparatus further comprises a means for clustering the pieces of music in the database into classes.

22. An apparatus according to any of claims 18 to 21, whereby a classification according to which the pieces of music are indexed uses feature extraction means, the means comprising:

(g) segmentation means for segmenting an entire piece of music over time into a plurality of windows;
(h) feature extraction means for extracting one or more features in each of said windows; and
(i) histogram determination means for arranging the features in histograms wherein the histograms are representative of the features over the entire piece of music.

23. An apparatus method according to claim 22, whereby a first feature extracted in step (h) is at least one tempo extracted from a digitised music signal, and wherein the feature extraction means comprise:

(ha) segmentation means for segmenting the music signal into a plurality of windows;
(hb) energy determination means for determining values indicative of the energy in each window;
(hc) peak location determination means for locating the peaks of an energy signal which signal is derived from the energy values in each window;
(hd) onset signal generation means for generating an onset signal comprising pulses, where pulse peaks

substantially coincide with the peaks of the energy signal;
(he) a plurality of comb filter means for filtering the onset signal wherein the plurality of comb filter means have resonant frequencies located according to frequencies derived from the window segmentation;
(hf) energy accumulation means for accumulating an energy in each filter process over a duration of the music signal; and
(hg) identification means for identifying the filter processes having Nth highest energies wherein resonant frequencies of the identified processes are representative of at least one tempo in the music signal.

24. An apparatus according to claim 23, whereby the energy determination means in sub-step (hb) comprise:

(hba) transform determination means for determining transform components for the music signal in each window; and
(hbb) addition means for adding amplitudes of the components in each window to form a component sum, said component sum being indicative of energy in a window.

25. An apparatus according to claim 23 or 24, wherein the apparatus further comprises low pass filtering means for low pass filtering the energy signal after locating the peaks of an energy signal in sub-step (hc) and prior to forming the onset signal generated in sub-step (hd).

26. An apparatus according to claim 23, 24 or 25, wherein the onset signal generation means in sub-step (hd) comprise:

(hda) differentiating means for differentiating the energy signal; and
(hdb) rectification means for half-wave rectifying the differentiated signal to form the onset signal.

27. An apparatus according to claim 23, 24 or 25, wherein the onset signal generation means in sub-step (hd) comprise:

(hdc) sampling means for sampling the energy signal;
(hdd) comparator means for comparing consecutive samples to determine a positive peak; and
(hde) pulse generation means for generating a single pulse when each positive peak is detected.

28. An apparatus according to any of claims 23 to 27, wherein the comb filter means resonant frequencies span a frequency range substantially between 1Hz and 4Hz.

29. An apparatus according to claim 22, whereby a second feature extracted in step (h) is a percussivity of a signal, and wherein the feature extraction means comprise:

(hh) segmentation means for segmenting the signal into a plurality of windows, and for each window;
(hi) filtering means for filtering by a plurality of filters;
(hj) filter output determination means for determining an output for each filter;
(hk) function determination means for determining a function of the filter output values;
(hl) gradient determination means for determining a gradient for the linear function; and
(hm) percussivity determination means for determining a percussivity as a function of the gradient.

30. An apparatus according to claim 29, wherein the segmentation means in sub-step (hh) comprise:

(hha) selection means for selecting a window width;
(hhb) overlap determination means for selecting a window overlap extent: and
(hhc) segmentation means for segmenting the signal into windows each window having the selected window width and the windows overlapping each other to the selected overlap extent.

31. An apparatus according to claim 29 or 30, wherein the filtering means in sub-step (hi) are comb filters.

32. An apparatus according to claim 29, 30 or 31, whereby the gradient determination means in sub-step (hl) comprises means for determining a straight line of best fit to the linear function.

33. An apparatus according to any of claims 29 to 32, wherein the percussivity determination means in sub-step (hm) consolidate the percussivity for each window into a histogram.

34. A computer readable medium incorporating a computer program product for querying a music database, which contains a plurality of pieces of music wherein the pieces are indexed according to one or more parameters said computer program product comprising:

(a) a request means for forming a request which specifies one or more pieces of music and/or associated parameters and one or more conditional expressions;
(b) a parameter determination means for determining associated parameters for the specified pieces of music if the parameters have not been specified;
(c) a comparison means for comparing the specified parameters and corresponding parameters associated with other pieces of music in the database;
(d) a distance determination means for calculating a distance based on the comparisons; and
(e) a determination means for identifying pieces of music which are at distances from the specified pieces of music as to satisfy the conditional expressions.

35. A computer readable medium according to claim 34 said computer program product comprising:

(f) an output means for outputting the identified pieces of music and/or the names of said pieces.

36. A computer readable medium according to claim 34 or 35, whereby the distance determination means (d) comprises:

(da) a distance determination means for a loudness, a percussivity, and a sharpness; and
(db) a sorting means for a tempo.

37. A computer readable medium according to claim 34, 35 or 36, whereby the computer program product relating to the comparison means in (c) comprises a means for clustering the pieces of music in the database into classes.

38. A computer readable medium according to any of claims 34 to 37, whereby a classification according to which the pieces of music are indexed uses feature extraction means, said computer program product comprising:

(g) segmentation means for segmenting an entire piece of music over time into a plurality of windows;
(h) feature extraction means for extracting one or more features in each of said windows; and
(i) histogram determination means for arranging the features in histograms wherein the histograms are representative of the features over the entire piece of music.

39. A computer readable medium method according to claim 38, whereby a first feature extracted in step (h) is at least one tempo extracted from a digitised music signal, and wherein said computer program product comprising:

(ha) segmentation means for segmenting the music signal into a plurality of windows;
(hb) energy determination means for determining values indicative of the energy in each window;
(hc) peak location determination means for locating the peaks of an energy signal which signal is derived from the energy values in each window;
(hd) onset signal generation means for generating an onset signal comprising pulses, where pulse peaks substantially coincide with the peaks of the energy signal;
(he) a plurality of comb filter means for filtering the onset signal wherein the plurality of comb filter means have resonant frequencies located according to frequencies derived from the window segmentation;
(hf) energy accumulation means for accumulating an energy in each filter process over a duration of the music signal; and
(hg) identification means for identifying the filter processes having Nth highest energies wherein resonant frequencies of the identified processes are representative of at least one tempo in the music signal.

40. A computer readable medium according to claim 39, whereby said computer program product relating to the energy determination means in sub-step (hb) comprises:

(hba) transform determination means for determining transform components for the music signal in each window; and
(hbb) addition means for adding amplitudes of the components in each window to form a component sum, said component sum being indicative of energy in a window.

41. A computer readable medium according to claim 39 or 40, said computer program product further comprising low pass filtering means for low pass filtering the energy signal after locating the peaks of an energy signal in sub-step (hc) and prior to forming the onset signal generated in sub-step (hd).

42. A computer readable medium according to claim 39, 40 or 41, wherein said computer program product relating to the onset signal generation means in sub-step (hd) comprises:

(hda) differentiating means for differentiating the energy signal; and
(hdb) rectification means for half-wave rectifying the differentiated signal to form the onset signal.

43. A computer readable medium according to claim 39, 40 or 41, wherein said computer program product relating to the onset signal generation means in sub-step (hd) comprises:

(hdc) sampling means for sampling the energy signal;
(hdd) comparator means for comparing consecutive samples to determine a positive peak; and
(hde) pulse generation means for generating a single pulse when each positive peak is detected.

44. A computer readable medium according to any of claims 39 to 43, said computer program product relating to the filter means resonant frequencies spaning a frequency range substantially between 1Hz and 4Hz.

45. A computer readable medium according to claim 38, whereby a second feature extracted in step (h) is a percussivity of a signal, and wherein said computer program product relating to the feature extraction means comprise:

(hh) segmentation means for segmenting the signal into a plurality of windows, and for each window;
(hi) filtering means for filtering by a plurality of filters;
(hj) filter output determination means for determining an output for each filter;
(hk) function determination means for determining a function of the filter output values;
(hl) gradient determination means for determining a gradient for the linear function; and
(hm) percussivity determination means for determining a percussivity as a function of the gradient.

46. A computer readable medium according to claim 45, wherein said computer program product relating to the segmentation means in sub-step (hh) comprises:

(hha) selection means for selecting a window width;
(hhb) overlap determination means for selecting a window overlap extent; and
(hhc) segmentation means for segmenting the signal into windows each window having the selected window width and the windows overlapping each other to the selected overlap extent.

47. A computer readable medium according to claim 45 or 46, wherein the said computer program product relating to filtering means in sub-step (hi) are comb filters.

48. A computer readable medium according to claim 45, 46 or 47, whereby said computer program product relating to the gradient determination means in sub-step (hl) comprises means for determining a straight line of best fit to the linear function.

49. A computer readable medium according to any of claims 45 to 48, wherein said computer program product relating to the percussivity determination means in sub-step (hm) consolidates the percussivity for each window into a histogram.

50. A method for querying a music database substantially as described herein with reference to any one of the embodiments, as that embodiment is shown in the accompanying drawings.

51. An apparatus for querying a music database substantially as described herein with reference to any one of the embodiments, as that embodiment is shown in the accompanying drawings.

52. A computer readable medium for querying a music database substantially as described herein with reference to any one of the embodiments, as that embodiment is shown in the accompanying drawings

53. A method for querying a music database which contains a plurality of pieces of music, the method comprising the steps of:

receiving a user request for one or more pieces of music from the database;
generating one or more features representative of the style of the requested one or more pieces of music;
determining one or more features representative of the style of each piece of music in said database, if those features have not already been determined,
comparing the one or more features representative of the style of the pieces of music in the database with the one or more features representative of the one or more pieces of music requested by said request; and
identifying one or more piece of music from said database in response to said comparison step.

54. A signal carrying instructions for configuring a programmable processing device as an apparatus for querying a music database according to any of claims 18 to 33.

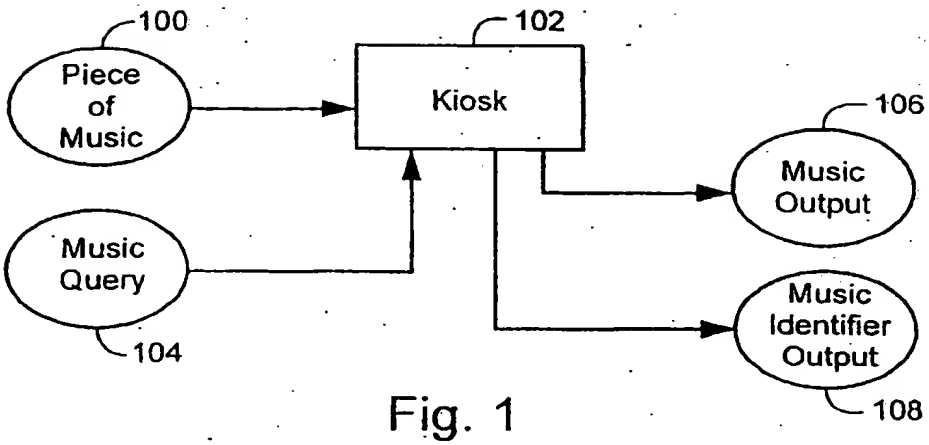
100
Piece of Music
102
Kiosk
106
Music Output
Music Query
104
Music Identifier Output
108

Fig. 1

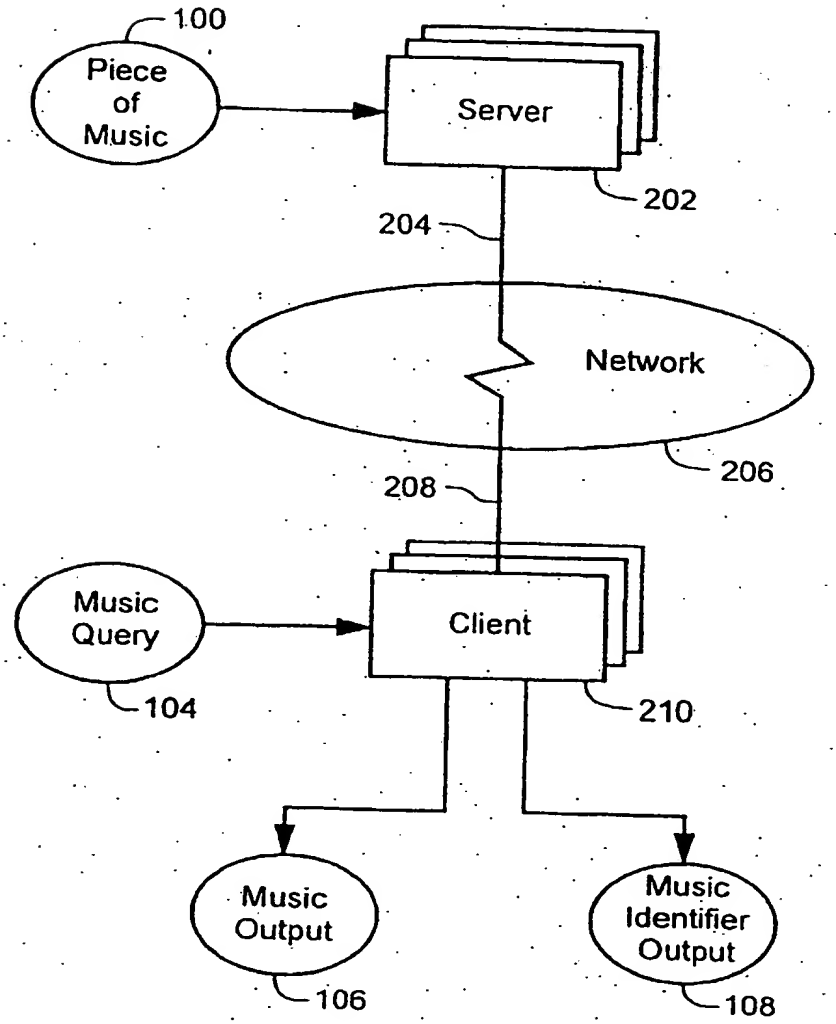
100
Piece of Music
Server
202
204
Network
206
208
Music Query
Client
104
210
Music Output
106
Music Identifier Output
108

Fig. 2

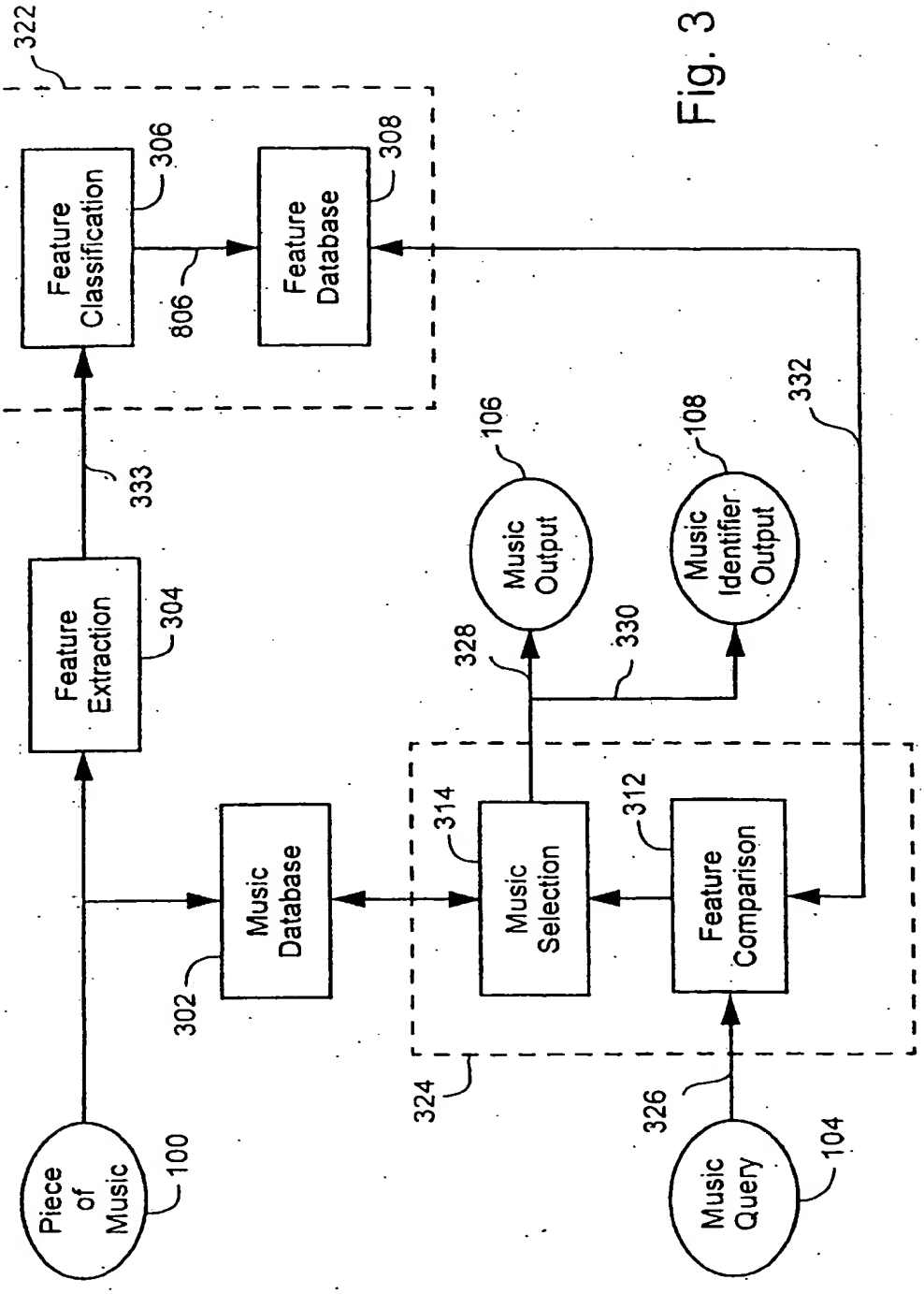
322
Feature Classification
306
Feature Database
308
806
333
Feature Extraction
304
Music Database
302
Piece of Music
100
314
Music Selection
312
Feature Comparison
324
326
Music Query
104
106
Music Output
108
Music Identifier Output
328
330
332

Fig. 3

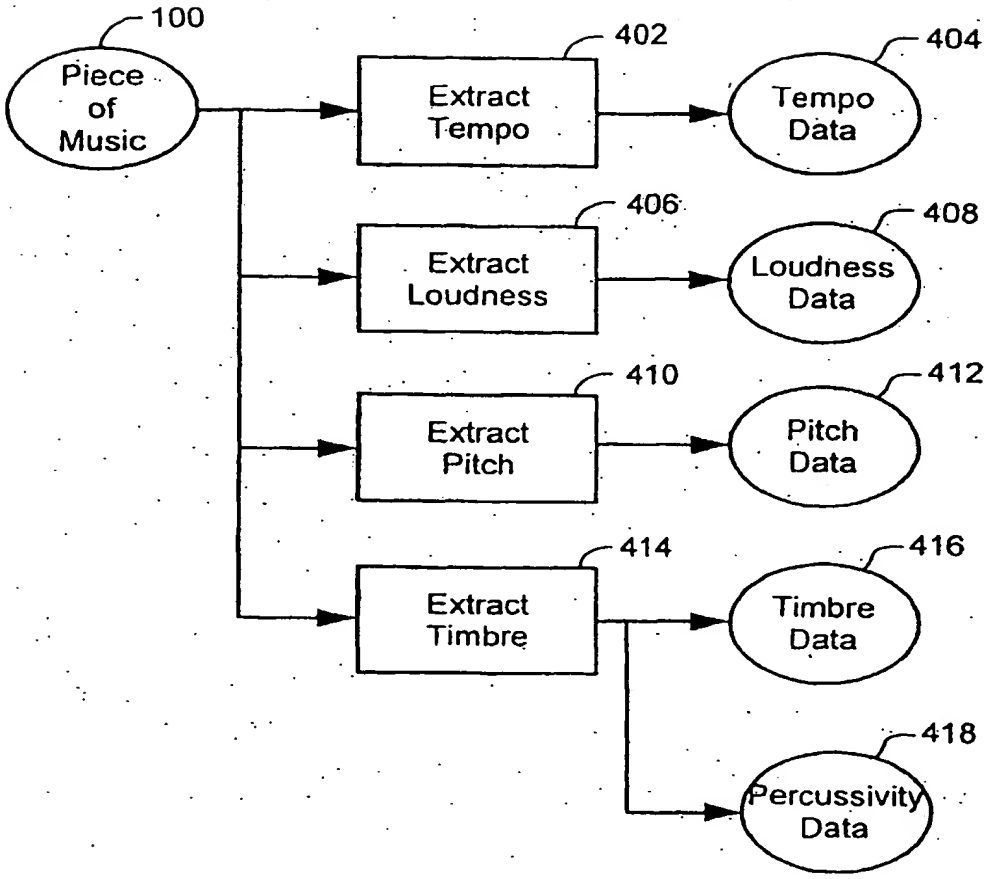
100
Piece of Music
402
Extract Tempo
404
Tempo Data
406
Extract Loudness
408
Loudness Data
410
Extract Pitch
412
Pitch Data
414
Extract Timbre
416
Timbre Data
418
Percussivity Data

Fig. 4

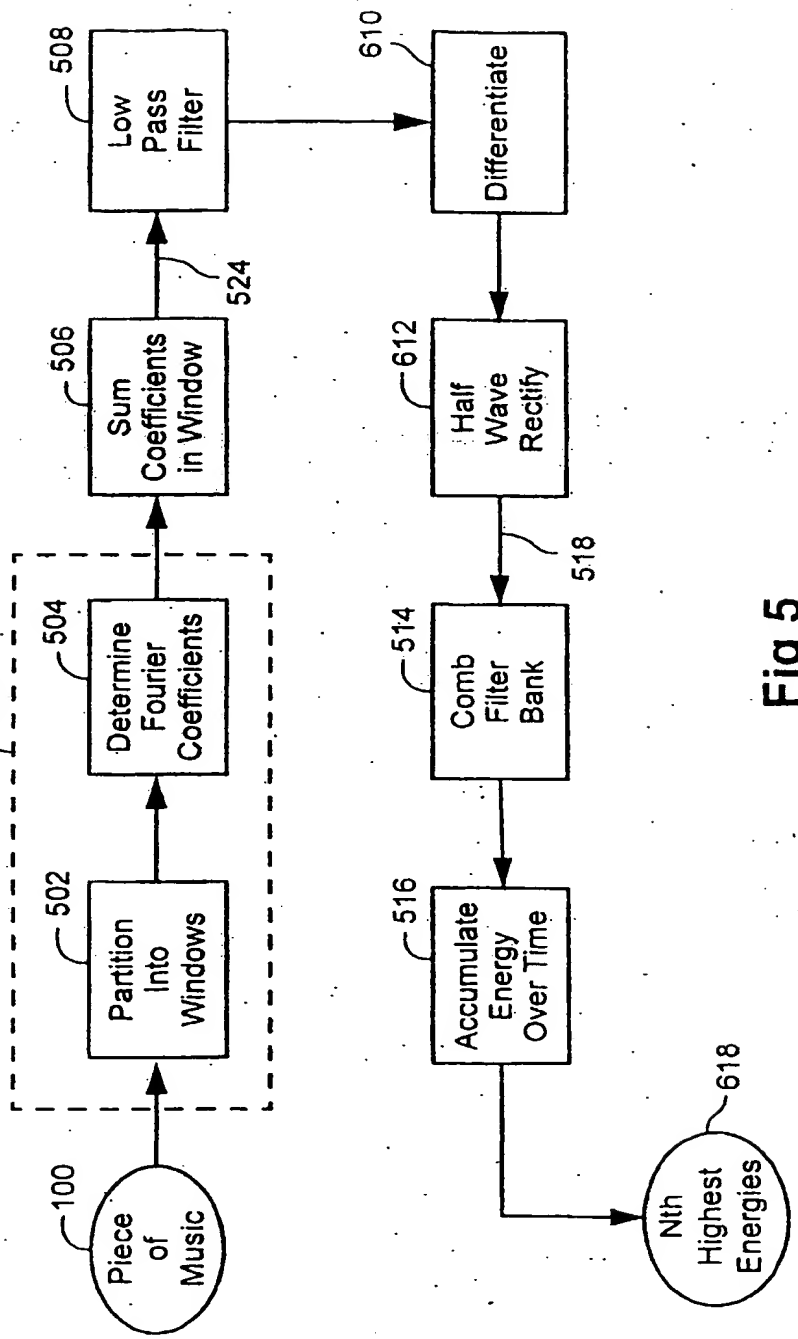
100
Piece of Music
522
502
Partition Into Windows
504
Determine Fourier Coefficients
506
Sum Coefficients in Window
524
508
Low Pass Filter
610
Differentiate
612
Half Wave Rectify
518
514
Comb Filter Bank
516
Accumulate Energy Over Time
618
Nth Highest Energies

Fig 5

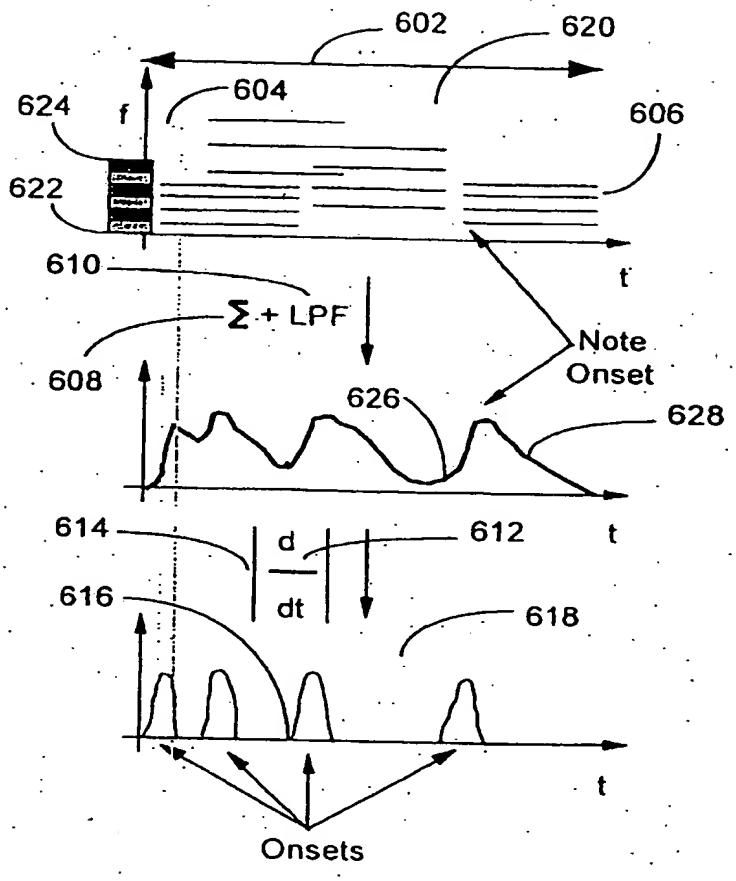
602
620
624
f
604
606
622
610
t
Σ + LPF
Note
Onset
608
626
628
614
d
dt
612
t
616
618
t
Onsets

Fig. 6

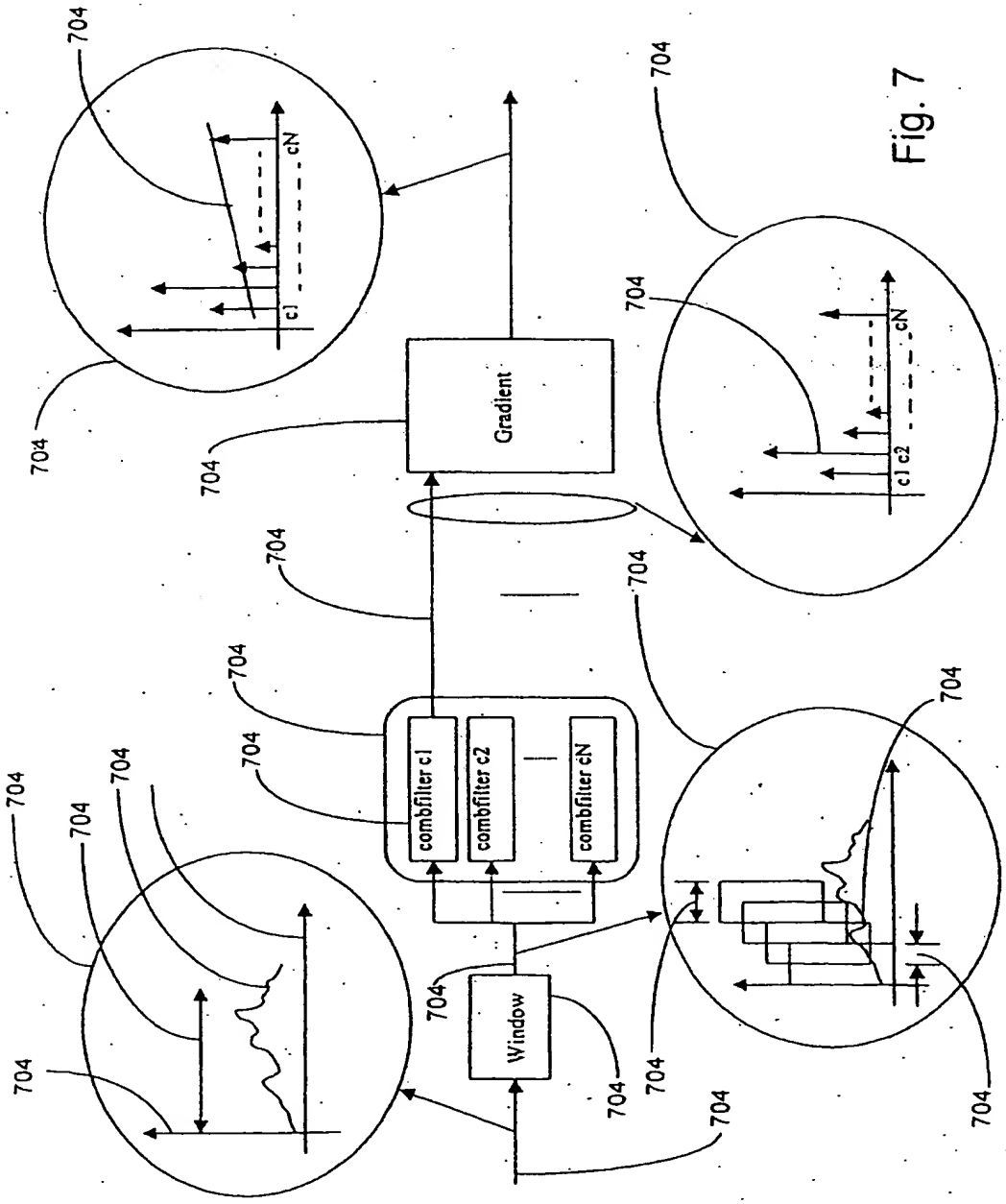
704
c1
cN
Gradient
c1 c2
combfilter c1
combfilter c2
combfilter cN
Window

Fig. 7

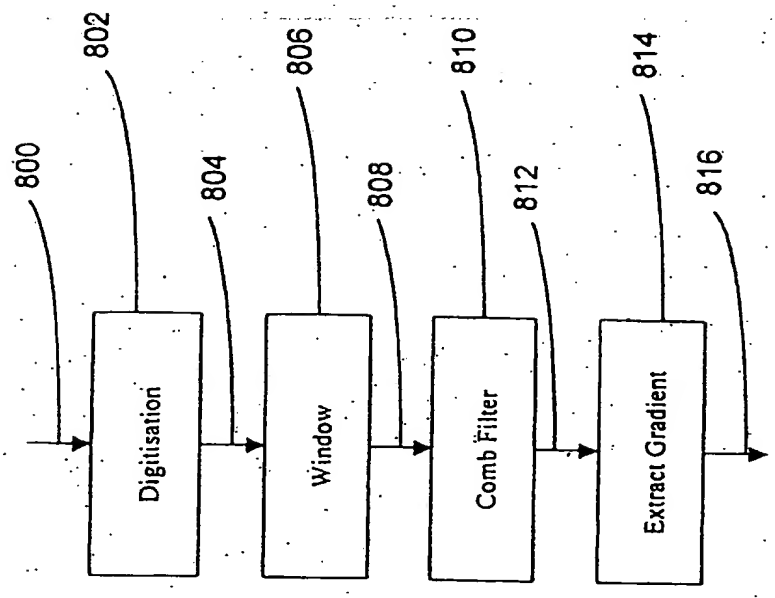
800
802
804
806
808
810
812
814
816
Digitisation
Window
Comb Filter
Extract Gradient

Fig. 8

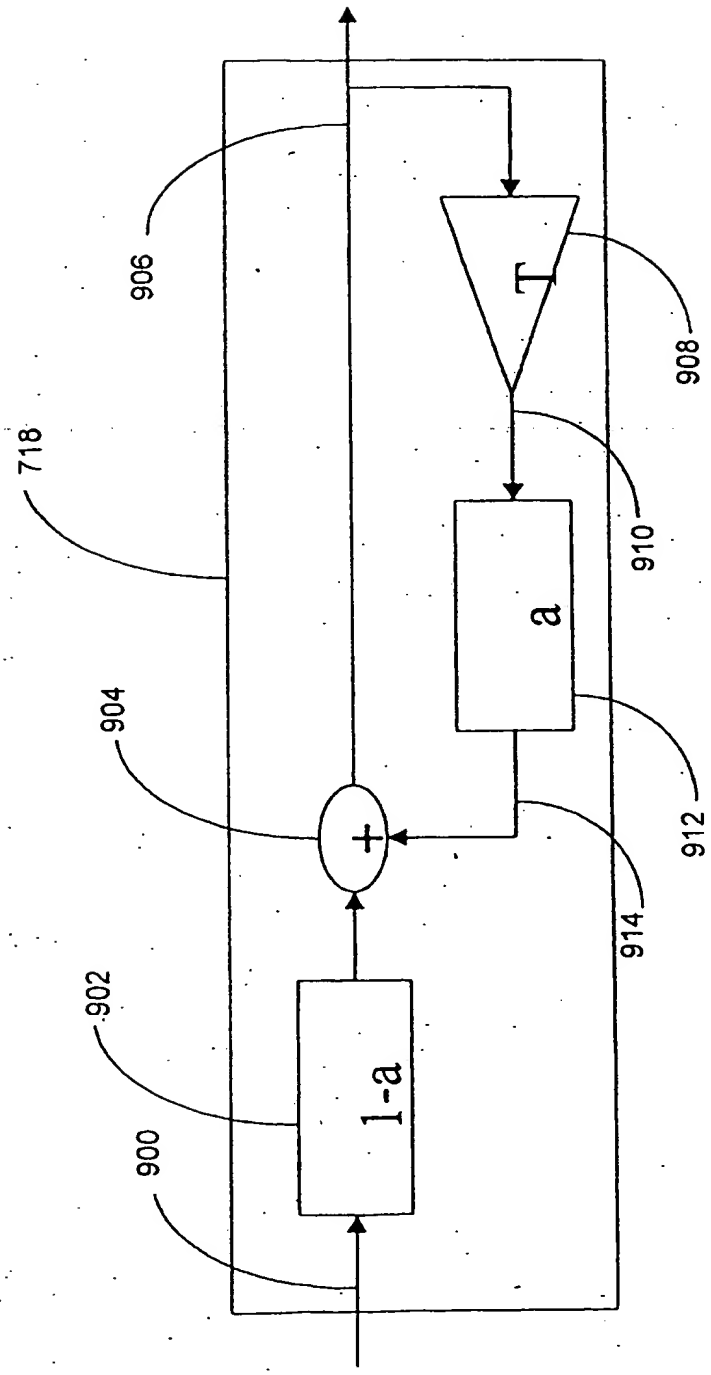


Fig. 9

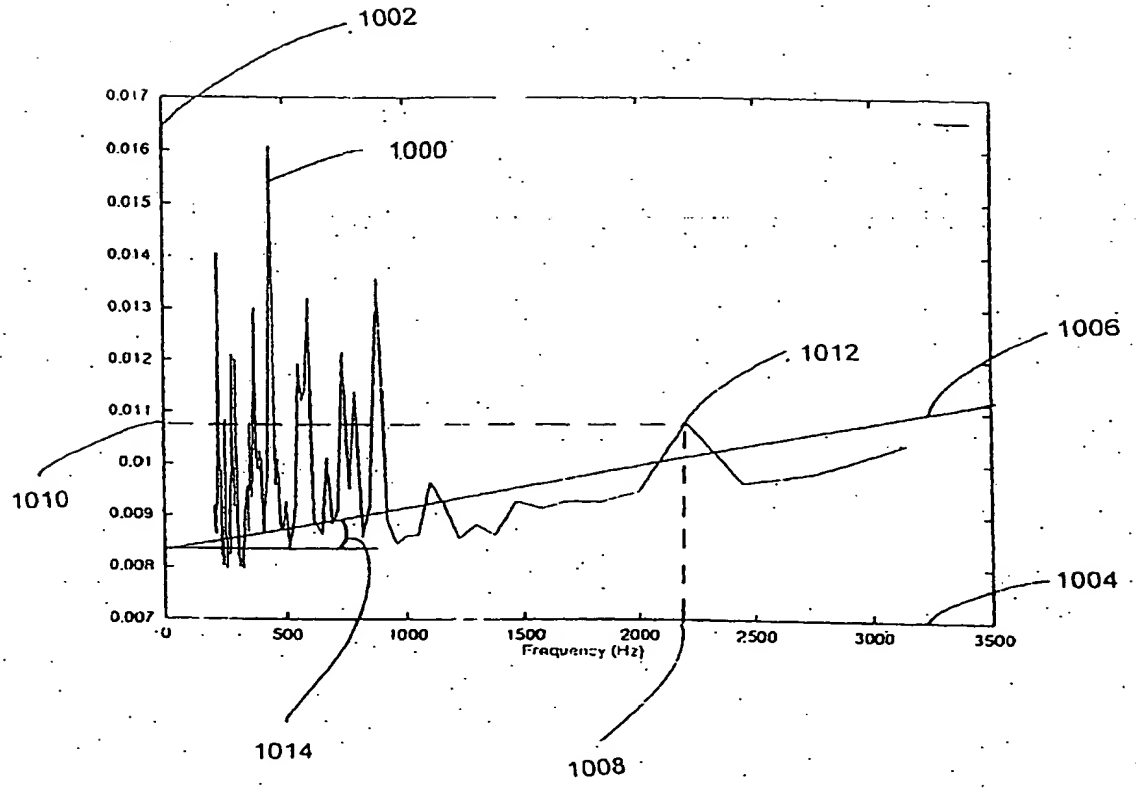
1002
1000
1012
1006
1010
1004
1014
1008
0.017
0.016
0.015
0.014
0.013
0.012
0.011
0.01
0.009
0.008
0.007
0
500
1000
1500
2000
2500
3000
3500
Frequency (Hz)

Fig. 10

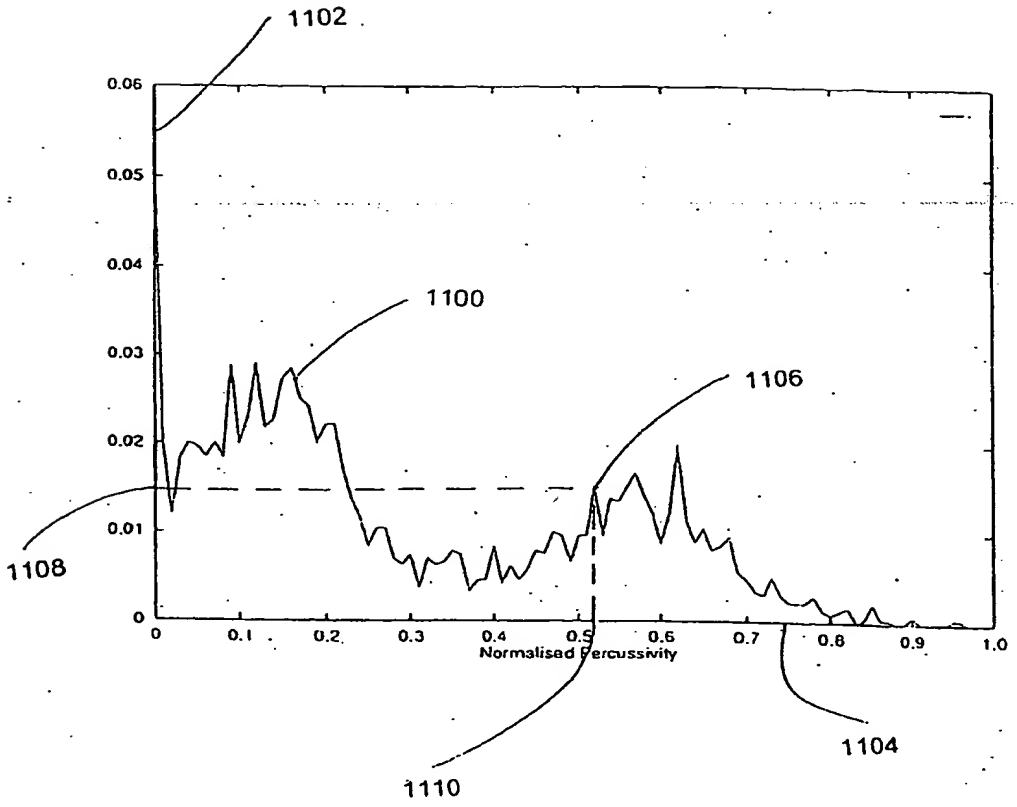
1102
0.06
0.05
0.04
0.03
0.02
0.01
0
1100
1106
1108
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Normalised Percussivity
1104
1110

Fig. 11

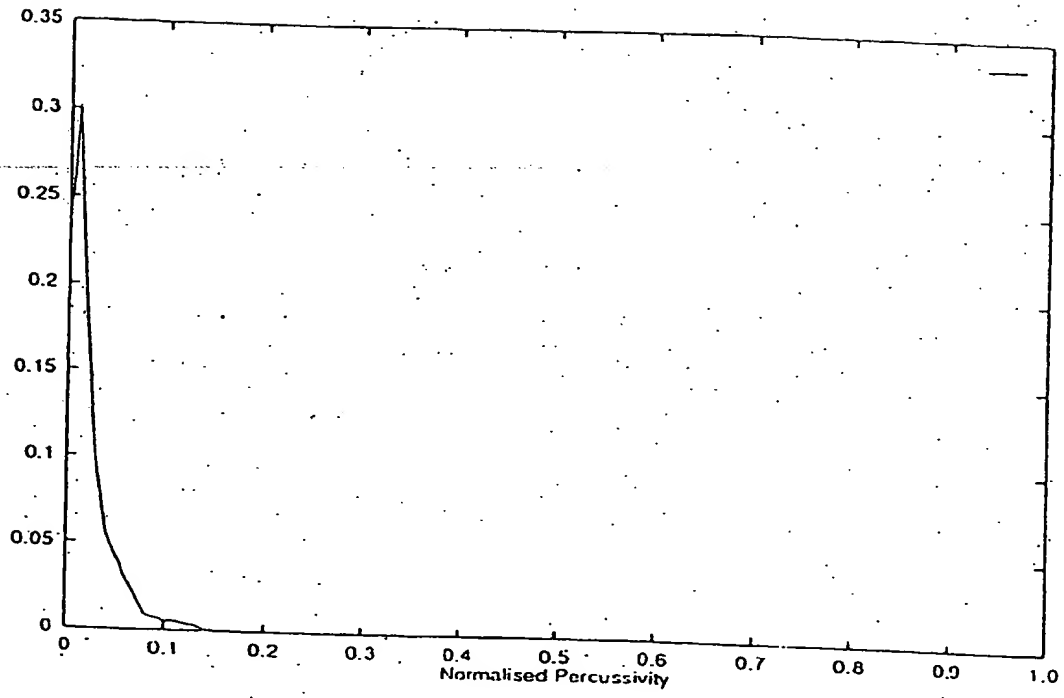
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Normalised Percussivity

Fig. 12

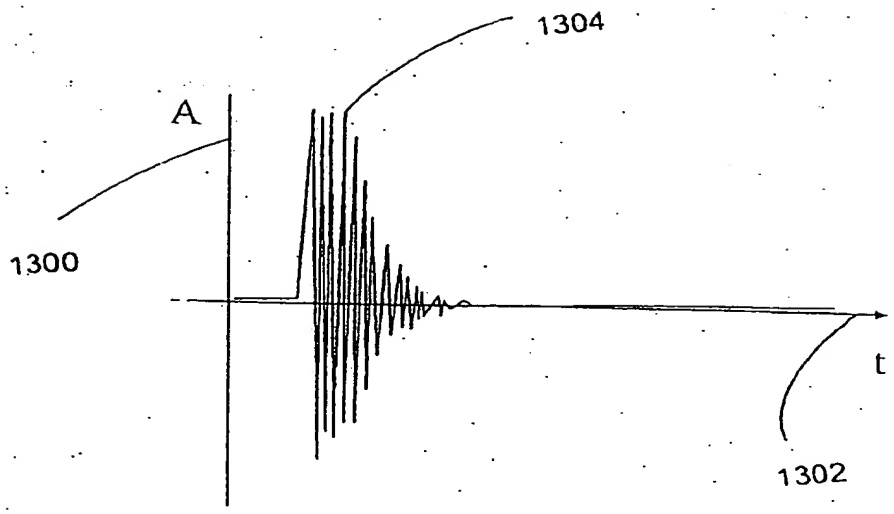


Fig. 13

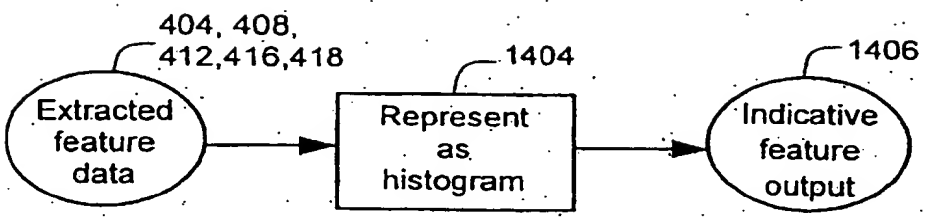
404, 408,
412,416,418
1404
1406
Extracted feature data
Represent as histogram
Indicative feature output

Fig. 14

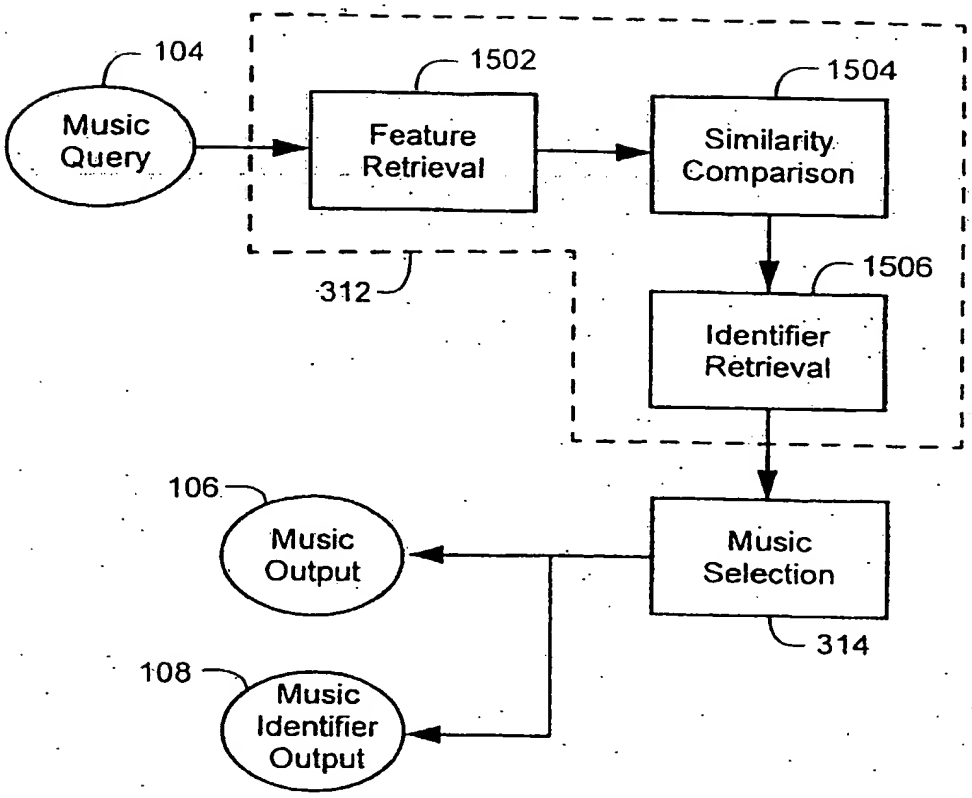
104
Music Query
1502
Feature Retrieval
1504
Similarity Comparison
312
1506
Identifier Retrieval
106
Music Output
Music Selection
314
108
Music Identifier Output

Fig. 15

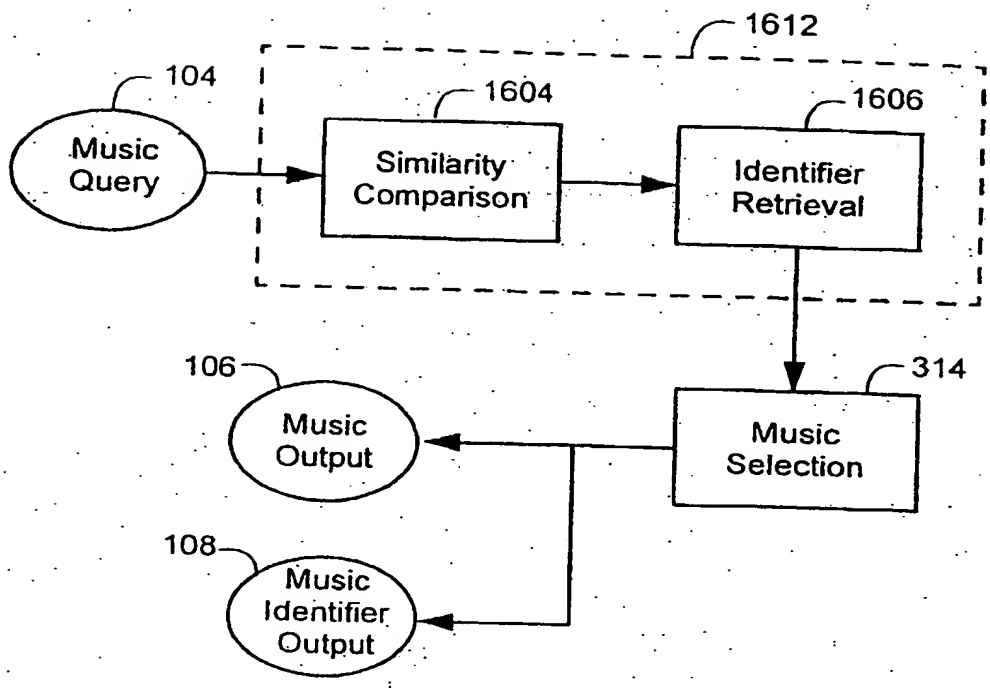
1612
104
Music Query
1604
Similarity Comparison
1606
Identifier Retrieval
106
Music Output
314
Music Selection
108
Music Identifier Output

Fig. 16

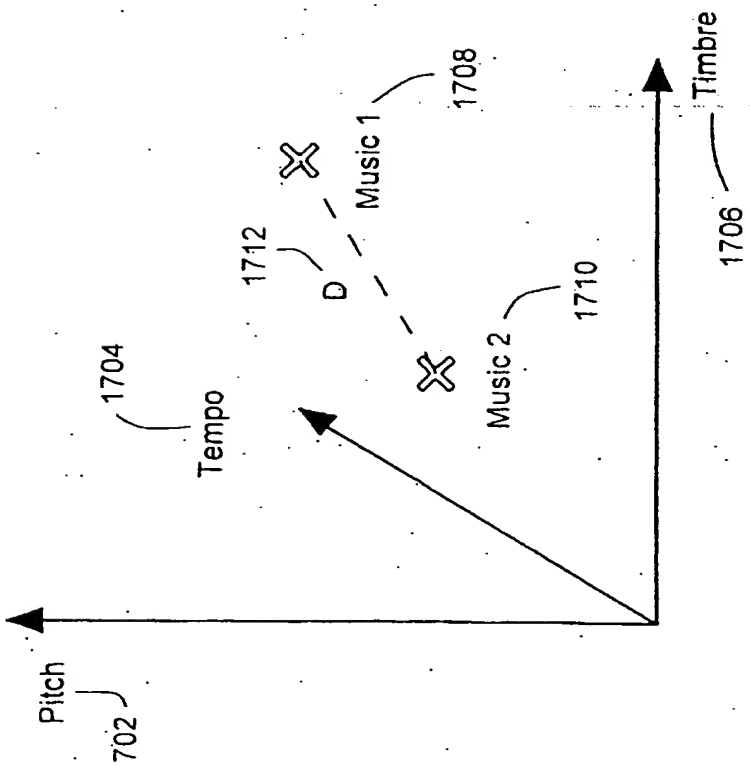
Pitch
702
Tempo
1704
Timbre
1706
Music 1
1708
Music 2
1710
D
1712

Fig. 17

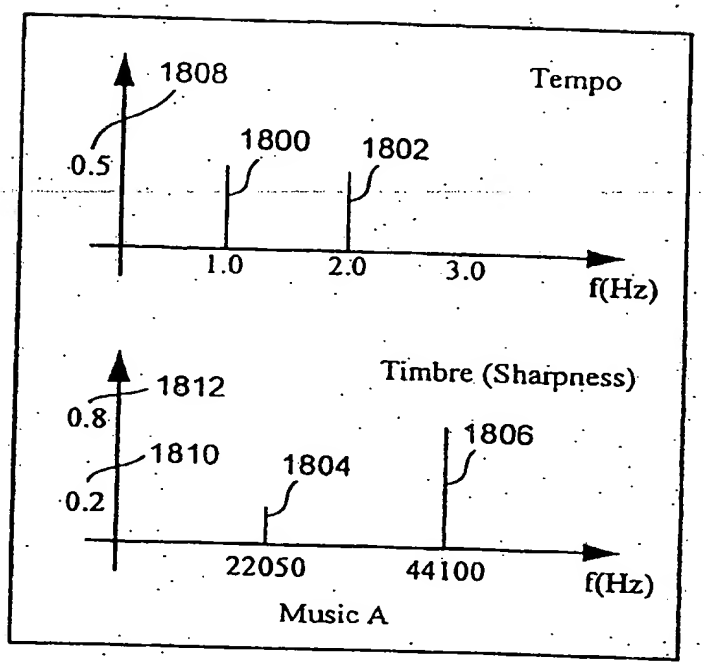
1808
Tempo
0.5
1800
1802
1.0
2.0
3.0
f(Hz)
1812
Timbre (Sharpness)
0.8
1810
0.2
1804
1806
22050
44100
f(Hz)
Music A

Fig. 18

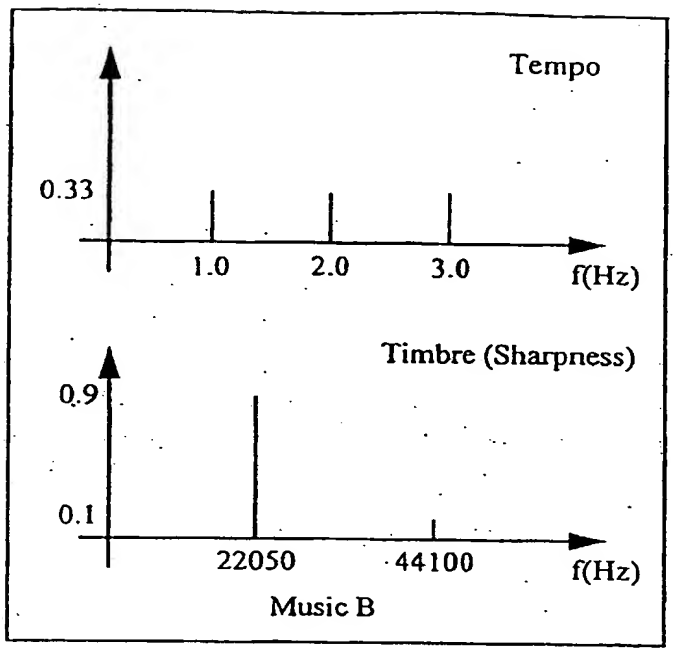
Tempo
0.33
1.0
2.0
3.0
f(Hz)
Timbre (Sharpness)
0.9
0.1
22050
44100
f(Hz)
Music B

Fig. 19

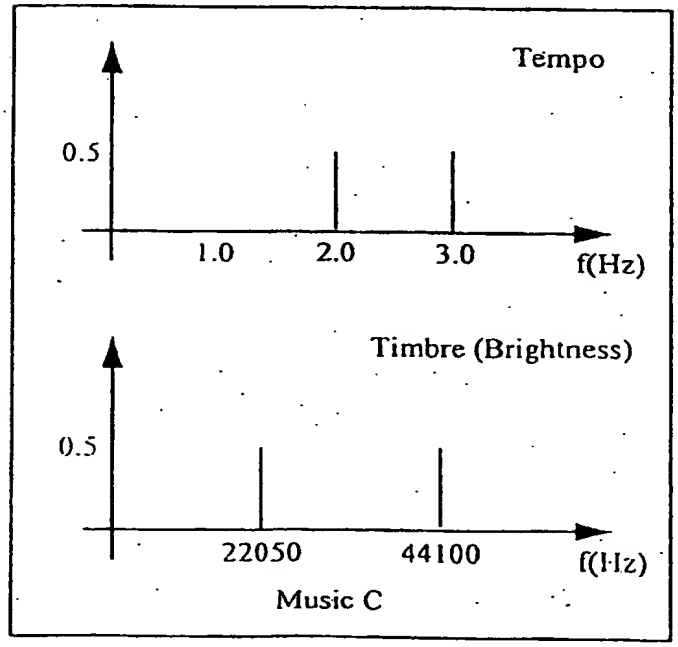
Tempo
0.5
1.0
2.0
3.0
f(Hz)
Timbre (Brightness)
0.5
22050
44100
f(Hz)
Music C

Fig. 20

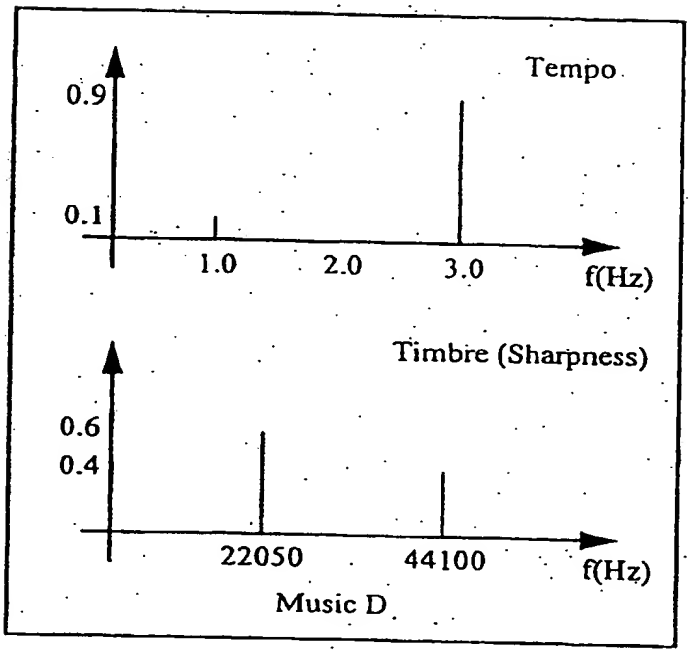
Tempo
0.9
0.1
1.0
2.0
3.0
f(Hz)
Timbre (Sharpness)
0.6
0.4
22050
44100
f(Hz)
Music D

Fig. 21

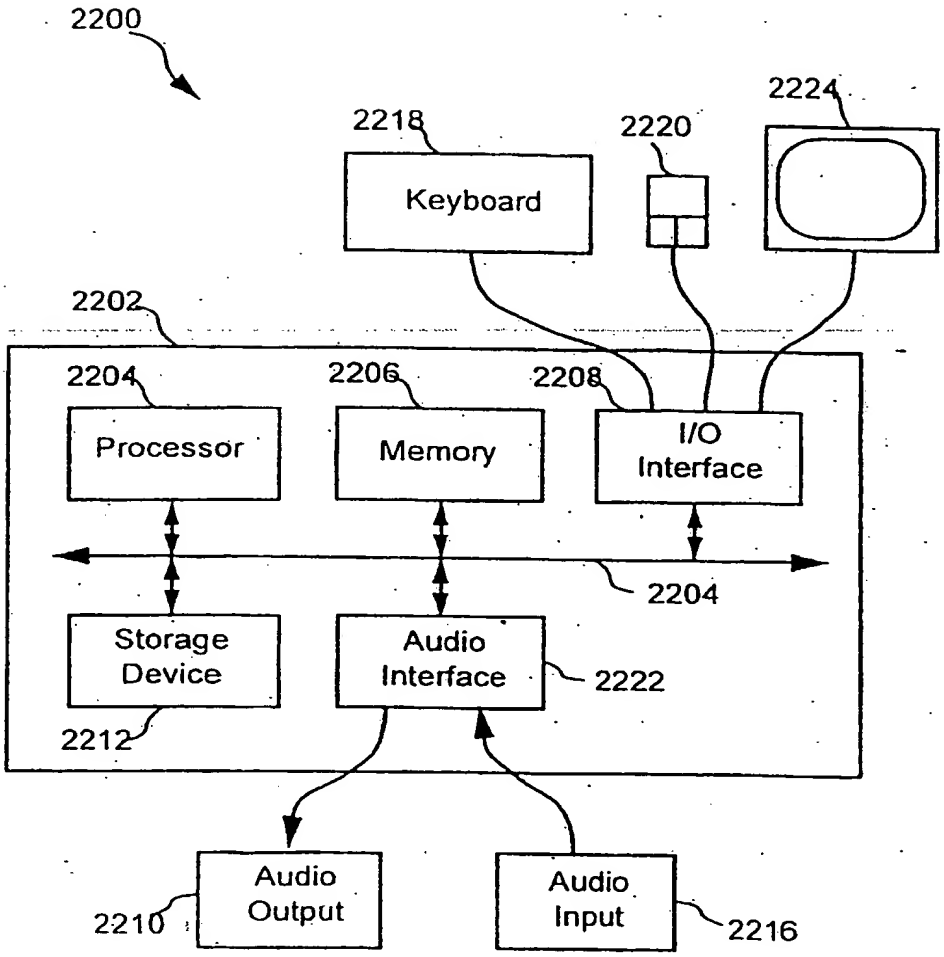
2200
2218
Keyboard
2220
2224
2202
2204
Processor
2206
Memory
2208
I/O Interface
2204
Storage Device
2212
Audio Interface
2222
Audio Output
2210
Audio Input
2216

Fig 22

(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 0 955 592 A3

# (12) EUROPEAN PATENT APPLICATION

(88) Date of publication A3:
31.01.2001 Bulletin 2001/05

(51) Int Cl.7: G06F 17/30, G10H 1/00

(43) Date of publication A2:
10.11.1999 Bulletin 1999/45

(21) Application number: 99303432.1

(22) Date of filing: 30.04.1999

(84) Designated Contracting States:
AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE
Designated Extension States:
AL LT LV MK RO SI

(30) Priority: 07.05.1998 AU PP340598
07.05.1998 AU PP340898
07.05.1998 AU PP341098

(71) Applicant: CANON KABUSHIKI KAISHA
Tokyo (JP)

(72) Inventor: Yourlo, Zhenya
Roseville, New South Wales 2069 (AU)

(74) Representative:
Beresford, Keith Denis Lewis et al
BERESFORD & Co.
High Holborn
2-5 Warwick Court
London WC1R 5DJ (GB)

## (54) A system and method for querying a music database

(57) A system and method for querying a music database (302), the database containing a plurality of indexed pieces of music, where the query (104) is performed by forming a database request consisting of a conditional expression relating to the name and/or attributes of the desired piece of music. Associated parameters are derived from the database query, and compared with corresponding parameters for the other pieces of music in the database (302). A desired piece of music is determined by searching for a minimum distance between the database query parameters and those associated with the pieces of music in the database (302).

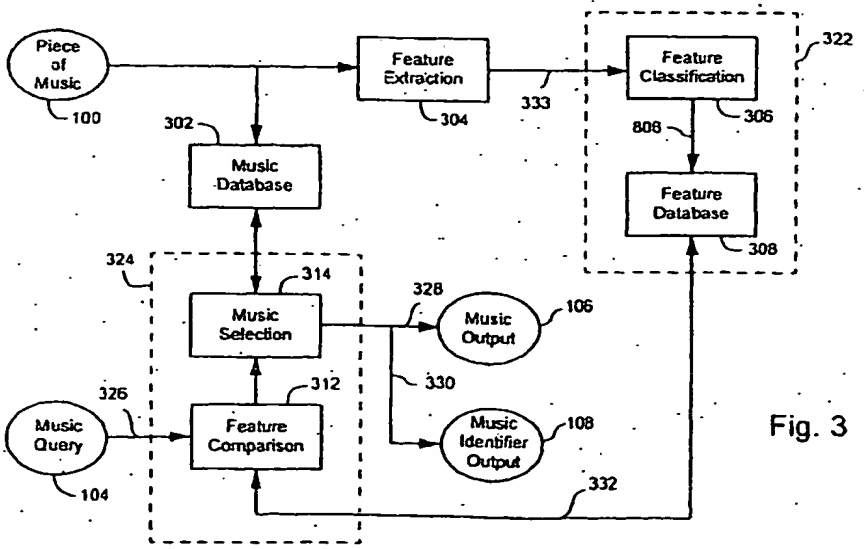


Fig. 3

EP 0 955 592 A3

European Patent Office

# EUROPEAN SEARCH REPORT

Application Number
EP 99 30 3432

## DOCUMENTS CONSIDERED TO BE RELEVANT

| Category | Citation of document with indication, where appropriate, of relevant passages | Relevant to claim | CLASSIFICATION OF THE APPLICATION (Int.Cl.6) |
|---|---|---|---|
| D,X | WOLD E ET AL: "Content-based classification, search, and retrieval of audio" IEEE MULTIMEDIA, FALL 1996, IEEE, USA, vol. 3, no. 3, pages 27-36, XP002154735 ISSN: 1070-986X | 1,2,4, 18,19, 21,34, 35,37, 50-52,54 | G06F17/30 G10H1/00 |
| A | * page 27, left-hand column, line 1 - page 28, right-hand column, line 41 * * page 30, left-hand column, line 29 - right-hand column, line 41 * * page 32, right-hand column, line 28 - page 34, right-hand column, line 2 * | 53 | |
| X | TA-CHUN CHOU ET AL: "Music databases: indexing techniques and implementation" PROCEEDINGS. INTERNATIONAL WORKSHOP ON MULTI-MEDIA DATABASE MANAGEMENT SYSTEMS (CAT. NO.96TB100064), PROCEEDINGS OF INTERNATIONAL WORKSHOP ON MULTIMEDIA DATABASE MANAGEMENT SYSTEMS, BLUE MOUNTAIN LAKE, NY, USA, 14-16 AUG. 1996, pages 46-53, XP002154736 1996, Los Alamitos, CA, USA, IEEE Comput. Soc. Press, USA ISBN: 0-8186-7469-5 | 1,18,34, 50-52,54 | TECHNICAL FIELDS SEARCHED (Int.Cl.6) G10H |
| A | * page 47, left-hand column, line 7 - right-hand column, line 11 * * page 52, left-hand column, paragraph 4; figure 4.1 * | 2-17, 19-33, 35-49,53 | |
| X | PATENT ABSTRACTS OF JAPAN vol. 1998, no. 03, 27 February 1998 (1998-02-27) & JP 09 293083 A (TOSHIBA CORP), 11 November 1997 (1997-11-11) * abstract * | 1,18,34, 50-52,54 | |

-/--

The present search report has been drawn up for all claims

| Place of search | Date of completion of the search | Examiner |
|---|---|---|
| THE HAGUE | 5 December 2000 | Fournier, C |

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)

European Patent Office

# EUROPEAN SEARCH REPORT

Application Number
EP 99 30 3432

## DOCUMENTS CONSIDERED TO BE RELEVANT

| Category | Citation of document with indication, where appropriate, of relevant passages | Relevant to claim | CLASSIFICATION OF THE APPLICATION (Int.Cl.6) |
|---|---|---|---|
| A | FOOTE J T: "Content-based retrieval of music and audio" MULTIMEDIA STORAGE AND ARCHIVING SYSTEMS II, DALLAS, TX, USA, 3-4 NOV. 1997, vol. 3229, pages 138-147, XP002154737 Proceedings of the SPIE - The International Society for Optical Engineering, 1997, SPIE-Int. Soc. Opt. Eng, USA ISSN: 0277-786X * the whole document * | 1-54 | |
| A | R. GONZALEZ & K. MELIH: "Content Based Retrieval of Audio" PROCEEDINGS OF THE AUSTRALIAN TELECOMMUNICATION NETWORKS &, 'Online! 1996, XP002154738 Retrieved from the Internet: <URL:http://www.int.gu.edu.au/MICTR/papers/gon96a.ps> 'retrieved on 2000-12-05! * the whole document * | 1-54 | |
| | | | TECHNICAL FIELDS SEARCHED (Int.Cl.6) |

The present search report has been drawn up for all claims

| Place of search | Date of completion of the search | Examiner |
|---|---|---|
| THE HAGUE | 5 December 2000 | Fournier, C |

CATEGORY OF CITED DOCUMENTS

X: particularly relevant if taken alone
Y: particularly relevant if combined with another document of the same category
A: technological background
O: non-written disclosure
P: intermediate document

T: theory or principle underlying the invention
E: earlier patent document, but published on, or after the filing date
D: document cited in the application
L: document cited for other reasons

&: member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)

## ANNEX TO THE EUROPEAN SEARCH REPORT ON EUROPEAN PATENT APPLICATION NO. EP 99 30 3432

This annex lists the patent family membersrelating to the patent documents cited in the above-mentioned European search report. The members are as contained in the European Patent Office EDP file on
The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

05-12-2000

| Patent document cited in search report | Publication date | Patent family member(s) | Publication date |
|---|---|---|---|
| JP 09293083 A | 11-11-1997 | NONE | |

EPO FORM P0459

For more details about this annex : see Official Journal of the European Patent Office, No. 12/82